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

Brief Measure Information

NQF #: 0553

Measure Title: Care for Older Adults (COA) – Medication Review

Measure Steward: National Committee for Quality Assurance

Brief Description of Measure: Percentage of adults 65 years and older who had a medication review during the measurement year. A medication review is a review of all a patient's medications, including prescription medications, over-the-counter (OTC) medications and herbal or supplemental therapies by a prescribing practitioner or clinical pharmacist.

Developer Rationale: Medication review can be a useful tool to reduce medication related problems (Christensen & Lundh 2012). The process of reviewing a patient's medication list reduces the risk of adverse drug interactions being overlooked and helps physicians minimize the duplication and complexity of the patient's medication regimen (Kallio et al. 2018). This in turn may increase patient adherence to the medication regimen and reduce hospital readmission rates. A recent systematic review found that medication review resulted in a decrease in the number of drug-related problems among patients (Huiskes et al. 2017).

Christensen, M., & Lundh, A. (2016). Medication review in hospitalised patients to reduce morbidity and mortality. Cochrane Database of Systematic Reviews, (2).

Huiskes, V. J. B., Burger, D. M., van den Ende, C. H. M., & van den Bemt, B. J. F. (2017). Effectiveness of medication review: a systematic review and meta-analysis of randomized controlled trials. BMC family practice, 18(1), 1-15.

Kallio, S. E., Kiiski, A., Airaksinen, M. S., Mäntylä, A. T., Kumpusalo-Vauhkonen, A. E., Järvensivu, T. P., & Pohjanoksa-Mäntylä, M. K. (2018). Community Pharmacists´ Contribution to Medication Reviews for Older Adults: A Systematic Review. Journal of the American Geriatrics Society, 66(8), 1613-1620.

Knight, E.L., J. Avorn. Quality indicators for appropriate medication use in vulnerable elders. Ann. Intern. Med. 2001. 703-10.

Numerator Statement: At least one medication review conducted by a prescribing practitioner or clinical pharmacist during the measurement year and the presence of a medication list in the medical record.

Denominator Statement: All patients 66 years and older as of the end (e.g., December 31) of the measurement year.

Denominator Exclusions: Exclude members who use hospice services.

Measure Type: Process

Data Source: Claims, Electronic Health Records, Paper Medical Records

Level of Analysis: Health Plan

Original Endorsement Date: Aug 05, 2009 **Most Recent Endorsement Date:** Aug 10, 2012

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 *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Summary of prior review in 2012

- In their last maintenance review in 2012, the developer provided a summary of the importance of medication review among the elderly, to reduce adverse drug events, one of the most common causes of mobidity and mortality in healthcare.

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:

- The developer provided a summary of the link between medication review and reduction in adverse drug events with improved health outcomes.
- The developer provided an Effectiveness of Medication Review: A Systematic Review and Meta-Analysis of Randomized Controlled Trials, that was published on BioMed Central (BMC) family practice, 18(1), 1-15., in 2017, by Victor Johan Bernard Huiskes, David Marinus Burger, Cornelia Helena Maria van den Ende, Bartholomeus Johannes Fredericus van den Bemt
    - In the review, here are the results: "Thirty-one RCTs were included in this systematic review (55% low risk of bias). A best evidence synthesis was conducted for 22 outcome measures. No effect of medication review was found on clinical outcomes (mortality, hospital admissions/healthcare use, the number of patients falling, physical and cognitive functioning), except a decrease in the number of falls per patient. However, in a sensitivity analysis using a

more stringent threshold for risk of bias, the conclusion for the effect on the number of falls changed to inconclusive. Furthermore no effect was found on quality of life and evidence was inconclusive about the effect on economical outcome measures. However, an effect was found on most drug-related problems: medication review resulted in a decrease in the number of drug-related problems, more changes in medication, more drugs with dosage decrease and a greater decrease or smaller increase of the number of drugs.
    - Conclusion: An isolated medication review during a short term intervention period has an effect on most drug-related outcomes, minimal effect on clinical outcomes and no effect on quality of life. No conclusion can be drawn about the effect on economical outcome measures. Therefore, it should be considered to stop performing cross-sectional medication reviews as standard care."
    - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the guideline.
        - The systematic review consists of 31 randiomized clinical trials.
        - The study used the Cochrane criteria and were evaluated according to their risk of bias, where 17 out of 31 studies (55%) met the criteria for low risk of bias. No grading was provided for this systematic review.
        - No harms were identified in the evidence.
- The developer provided a meta-analysis "Pharmacist-led medication review in community settings: An overview of systematic reviews" that was published in Research in Social and Administrative Pharmacy 13: 661-685 in 2017 by Jokanovic NJ, Tan ECK, Sudhakaran S, Kirkpatrick CM, Dooley MJ, Ryan-Atwood, TE, Bell JS.
    - Meta-analysis was performed in 12 systematic reviews examing the effect of pharmacist-led medication review conducted in community settings on clinical outcomes. Results suggested positive impacts on glycosylated hemoglobin (HbA1c), blood pressure, and cholesterol. No meta-analyses reported reduced mortality.

Exception to evidence
N/A

Questions for the Committee:

If the developer provided updated evidence for this measure:

- Based upon this 2017 review, is there a convincing relationship between medication review and outcomes that is sufficient to justify a quality measure?
- For structure, process, and intermediate outcome measures:
    - What is the relationship of this measure to patient outcomes?
    - How strong is the evidence for this relationship?
    - Is the evidence directly applicable to the process of care being measured?
    - If derived from patient report, does the target population value the measured process or structure and find it meaningful?

Guidance from the Evidence Algorithm

Process measure with systematic review (Box 3) → Summary of the QQC provided (Box 4) → Systematic review concludes moderate quality evidence.

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

The highest possible rating is "High" for Evidence

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer provided updated data, extracted from HEDIS, that represents the measurement years of 2014-2016, is stratified by year and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile.
    - Mean 87.2 (2014); Mean 84.1 (2015); Mean 88.0 (2016)
    - Number of Healh Plans (2014-2016): 404; 427; 504
    - STDEV: 16.1; 18.8; 14.0
    - MIN: 0.1; 0.0; 0.0
    - MAX: 100.0; 100.0; 100.0
    - Measurement Year: 2016; 2015; 2014
        - P10: 71.3; 57.9; 73.5
        - P25: 86.5; 80.3; 82.0
        - P50: 93.2; 91.0; 90.7
        - P75: 97.5; 96.6; 95.8
        - P90: 99.5; 98.9; 98.6

- Data from developer showed performance at 65% in Medicare patients, with considerable variability.

Disparities

- Per developer, this measure can be stratified by demographic variables, such as race, ethnicity or socioeconomic status, in order to assess the presence of health care disparities, in addition to being stratified by insurance type (Commercial, Medicaid, and Medicare) if the data are available in the plan. However the developer did not provide any disparities data on this measure.
- The developer recognized the presence of other studies that identify disparities in prescribing practices, but no studies were identified that could find disparities in rates of medication review.
- The developer cited a study by Trivedi et al 2018, which examined medication review and found no differences by race.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Since no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?

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

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

1a. Evidence
Comments:

**This process measure has some empirical data associated with it. However, there is little to suggest that medication review is affecting outcomes for patients.

**A process metric, logic: medication review reduces adverse drug events, new evidence.

**Okay.

**The evidence applies directly to the process measured. No additional evidence known other than what was submitted.

**Evidence shows questionable decrease in falls but does show a decrease in drug related problems. Pharmacist led med rec showed a decrease in HgA1C, BP and cholesterol.

**The evidence surrounding this measure should be discussed in the evaluation process. The 2017 review is worthy of discussion and the NQF measure review summarizes appropriate discussion points.

**Clear opportunity to improve.

**This is a flawed process measure. The shortcomings are: persons implementing the measure may not be qualified to deprescribe medication in patients with several comorbidities. Measure overlooks the fact that many in the age group addressed may be taking few medications, and therefore have no need for reconciliation. There is no provision for cases where the patient has a prescription and does not take it, or takes a lower dose of the prescription. In my opinion, this should be conversed to an outcome measure where the number of medications deprescribed is tallied. That is the outcome patients want.

**Found recent meta analysis interesting; actually quite limited relationship process to outcomes beyond med related harm.

**Evidence is moderate for this process measure, evidence of relationship to a few outcomes; new studies do not change evidence base.

**Developer provided updated information - moderate evidence for rationale and process measure.

**The primary source of evidence is a 2017 Systematic Literature Review that analyzes evidence of this measure on clinical outcomes (minimal effect), quality of life (no effective), economic outcomes (inconclusive) and drug-related outcome (positive effect to reduce drug related problems). Logic model: Adults age 65 and older >> medication review performed and medication list documented in the medical record >> reduction in adverse drug events >> improved health outcomes.

**This is a process measure 1 x month for age 65 and older. in a systematic review, no effect of medication review was found on clinical outcomes and inconclusive for falls using a sensitivity analysis

1b. Performance Gap
Comments:

**There is a performance gap demonstrated between health plans that were evaluated on this measure. Data on disparities were not included.

**Moderate, HEDIS MY 2014-2016 results showed continued gaps, HEDIS doesn't require data breakout by sub-population, but health plans can conduct their own disparities analysis.

**In the high 80 percent already.

**Performance data was provided and demonstrated variability. Disparity data not provided.

**Only 65% of medicare advantage patients have med rec. This is not stratified for disparities. One study showed no disparity by race.

**65% in medicare patients.

**There is a perfomrance gap-no data for disparities.

**Current performance data were provided. Disparities seemed small, but the measure fails to ensure deprescribing was appropriate.

**Gaps still exists across plans; no impressive improvements/gap reductions over time.

**High, 65% in medicare, wide variability.

**Demonstarted variability however no disparity data provided.

**Update analysis of HEDIS data on provider performance for 2014-2016 that found a 65% performance gap in Medicare patients and considerable variability. The developers did not provide disparities data on the measure.

**Gap exists in Medicare patients (65% performance)

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.

Composite measures only:

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

N/A

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

Evaluators: Patient Safety project team staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Link A (Project Team staff)

The developer conducted reliability testing,using beta-binomial model (signal to noise). In the previous 2012 submission, the signal to noise score for this measure, provided by the developer was calculated as 0.98712. The updated signal to noise score for this measure was calculated as 0.985 using 2016 data. The signal to noise result of 0.985 exceeds the 0.7 threshold.

The results from the Construct Validity testing, using a Pearson Correlation test by exploring whether the Care for Older Adults – Medication Review measure is correlated with the Care for Older Adults – Pain Assessment measure and found that organizations that perform well on the Care for Older Adults – Medication Review measures should also perform well on Care for Older Adults – Pain Assessment measureprovided by the developer, indicate that there is a strong, positive relationship between the Care for Older Adults – Medication Review measure and the Care for Older Adults – Pain Assessment measure. This relationship is statistically significant ($p<0.0001$). The developer also conducted Face Validity for this measure through various methods of data collection. Their multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity deemed the measure valid.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- Since this measure is tested in Medicare Advantage Special Needs Plans (SNPs), does that mean it can only be used for measurement in this population or is this patient population, which generally has more medical/support needs, a good representation of older adults?
- The staff 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 staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

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

Evaluation A: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0553

Measure Title: Care for Older Adults – Medication Review

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.**

   N/A

RELIABILITY: TESTING

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

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   ☐ **Yes** ☐ **No**

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

   Submission document: Testing attachment, section 2a2.2

   Appropriate beta-binomial model was used to calculate reliability using 2016 HEDIS data from 504 health plans (Medicare Advantage Special Needs Plans). Testing was conducted with data source and level of analysis indicated for this measure and conforms to NQF criteria and guidance.

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

   Submission document: Testing attachment, section 2a2.3

   A signal to noise analysis was conducted and score of 0.985 was calculated using an appropriate sample size.

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.**

    The testing results for this measure meet NQF's evaluation criteria for testing and demonstrates the measure data elements are repeatable. No concerns noted.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    N/A

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

    Submission document: Testing attachment, section 2b4.

    An inter-quartile range was calculated and shows there is a 11.0 percentage point gap between the 25th and 75th percentile, and that this difference is statistically significant (p<.0001). This gap represents an average of 37 more patients that have had their medications reviewed in high-performing plans compared to low-performing plans. No concerns with the ability to identify which meaningful differences.

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.

N/A – this measure has only one set of specifications.

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

Submission document: Testing attachment, section 2b6.

Per developer, this measure is collected with a complete sample.

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

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

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

16d.**Risk adjustment summary:** N/A

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

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

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

N/A

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**

18. **Method of establishing validity of the measure score:**

☒ **Face validity**

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

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

19. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

An appropriate Pearson correlation test was used to estimate the association between the Care for Older Adults – Medication Review measure and the Care for Older Adults – Pain Assessment measure. The two measures both focus on older adult patients and they both include a key component of assessment for this population. Both measures are reported by the same type of organization (Special Needs Plans), further ensuring a fair comparison.

A face validity assessment was also included as part of the submission.

20. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

The Pearson Correlation Coefficient was calculated to be 0.82. The correlation between the two measures is statistically significant at p<0.0001.

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒**Yes**

☐**No**

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

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

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

☐**Yes**

☐**No**

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

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

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

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

☐**Low** (NOTE: Should rate LOW if you believe that there 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.)

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

The construct validity and supporting face validity assessments included appropriate methods, yielded acceptable testing results, and conform to NQF criteria.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

25. **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**

N/A

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

N/A

ADDITIONAL RECOMMENDATIONS

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

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

2a1. Reliability – Specifications
Comments:

**I have no concerns about the reliability of the measure, and I believe it can be consistently implemented.

**Detailed specifications provided.

**Still relies on attestation.

**No concerns on reliability.

**Data elements clear.

**High reliability.

**Good reliability-no risk adjustment.

**The numerator should not include patients taking few medications. Deprescribing for them is not of much value. Measure can be consistently implemented because it is quite simple, and thereby, be of little benefit to the patient.

**Acceptable.

**High/moderate.

**No concerns for reliability.

**This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via NCQA's online data submission system. Signal to noise testing exceeded 0.985. No issues noted.

**high reliability

2a2. Reliability – Testing
Comments:

**High, new reliability score of 0.985 exceeds the 0.7 minimum required.

**Yes, too much reliance on attestation.

**No concerns on reliability.

**Signal to noise 0.985 well above 0.7.

**No.

**No.

**It's reliability is not optimal because it would be too easy to simply go through the motions of medication reconciliation, meet the requirements of the measure, and provide no benefit to the patient.

**No.

**No.

**I think the SNP group was a good group to test it in.

**Signal to noise testing. No concerns.

**No

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:

**No.

**Moderate, face validity and, Pearson Correlation Coefficient 0.82 with another metric.

**No.

**No concerns on testing results.

**Pearson correlation statistically significant.

**No.

**No.

**No.

**No.

**Moderate: Not sure how the researchers picked pain treatment as a validation variable

**None.

**Face and Empirical Validity testing reported. 0.82 correlation between Pain Assessment Measure and Medication Reconciliation. Construct validity and face validity conform with NQF requirements.

**No concerns, data collected from complete population.

**No.

**No concerns regarding validity of performance scores.

**No concerns.

**Minimal.

**No.

**No.

**No concerns.

**No.

**None.

**Testing completed in 2016. No missing data. No concerns with validity of the measure.

**moderate - no

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**I am not sure why medication review for hospice patients is an exclusion. Risk adjustment is not applicable.

**None, exclusions appropriate.

**Nothing new.

**No concerns over exclusions and Risk Adjustment is not applicable.

**No concerns.

**Appears appropriate.

**No risk adjustment-only exlcusion is hospice care.

**Excluding hospice patients may not be appropriate in some cases. Patients may be in hospice for a very long time and could benefit from medication reconciliation, but I agree that most do not.

**Unclear why hospice patients excluded.

**None.

**No concerns. All adults 66 and older. Excludes hospice patients.

**OK

**Medicare patients

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.

- Data elements for this measure can be collected through multiple data sources, such as: administrative data, electronic health data, and paper records
- Some data elements are in defined fields in electronic sources
- There is no actual charge/fee for inclusion of the measure
- This measure is not an eMeasure.

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?
- Is the data collection strategy ready to be put into operational use?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**No concerns about feasibility.

**High, administrative data and medical records.

**Wish it could also capture change as a result of review.

**High feasibility for effective application of measure.

**Administrative, electronic and paper routinely gathered during administration of care.

**Agree with high feasibility prelim rating.

**Can be either collected elctronically or on paper.

**None.

**No concerns.

**None.

**High feasibility.

**High rating.Data abstracted from administrative databases, electronic medical records.

**claims data, EHR and paper records - what is actually being used and to what extent - Burden of collection?

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**

Accountability program details

- CMS Star Ratings for Medicare Advantage Plans and Prescription Drug Plans
  - This measure is included in the composite Medicare Advantage Star Rating.
- Also public reported The HEALTHCARE EFFECTIVENESS AND DATA INFORMATION SET (HEDIS)
  - Includes measures for physicians, PPOs, and other organizations

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 Special Needs Plans that report HEDIS, and are notified of their performance when they submit their results. CMS reporting these rates for these plans, help Medicare Managed Care Organizations compare their performance to other like organizations.

Additional Feedback:

- Data results are reported annually, by CMS.
- The developer shares data regularly, via webinars, conferences and anualy reports and provides technical assistance on measures through their Policy Clarification Support System.

***Questions for the Committee**:*

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

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

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

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

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

Improvement results

- Data from developer showed performance at 65% in Medicare patients, with considerable variability.
    - 2014 Mean 87.2
    - 2015 Mean 84.1
    - 2016 Mean 88.0

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

- No unexpected findings were found during implementation.

Potential harms

- The developer indicated that no ptential harms were reported during the testing of this measure.

Additional Feedback:

- No unexpected benefits were reported.

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

4a1. Use - Accountability and Transparency

Comments:
**The measure is part of the CMS evaluation of Medicare Advantage plans, and is publicly available. It is not clear that health plans are using the measure to improve performance.
**High, HEDIS reported to NCQA for health plan accreditation and CMS for Star ratings.
**Yes.
**Adoption into public reporting without known issue or concern.
**Used in STAR ratings and HEDIS to compare performance in managed medicare plans. Opportunity for feedback given.
**Initial approval 2009.
**Publically reported not cleatr how much progress is being made.
**Somewhat.
**Do not see meaningful improvements over time; measure many not "matter".
**No concerns.
**No issues and already reported.
**Public Reporting. CMS Star Ratings for Medicare Advantage Plans and Prescription Drug Plans. Quality Improvement (external benchmarking to organizations) Healthcare Effectiveness Data and Information Set.
**Feedback at the plan level - not sure

4b1. Usability – Improvement
Comments:

** Medication review is a component of overall medication safety, and the measure has been used to change dosages and/or medications used. This contributes to quality and safety for patients.There are not unintended consequences that outweigh the benefits of the measure.

**No harms reported.

**Reasonable.

**Strong performance results demonstrate health plan commitment and support to minimize medication related errors and mishaps. No unintended consequences appear to be observed.

**Demonstrated improvement over time. No harms.

**Minimal apparent unintended consequences.

**Med reconciliation especially in this age group very important.

**As noted, consideration should be given to converting this to an outcome measure, either as the number of medications deprescribed or in better health for the patient.

** No concerns.

**Few unintended consequences.

** Usable and no unexepcted consqequences.

** Reported publicly and currently in use for accountability. Data results are reported annually by CMS. Measure is included in the composite Medicare Advantage Star Rating. Also public reported in HEDIS. • The developer shares data regularly, via webinars, conferences and annually reports and provides technical assistance on measures through their Policy Clarification Support System.

**Lack of effectiveness?

Criterion 5: Related and Competing Measures

Related or competing measures

- 0097 : Medication Reconciliation Post-Discharge
- 0419e : Documentation of Current Medications in the Medical Record
- 2456 : Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

- 2988 : Medication Reconciliation for Patients Receiving Care at Dialysis Facilities
- 3317 : Medication Reconciliation on Admission

Harmonization

The developer noted that this measure has been harmonized as much as possible and that measure 0553 s differs from other related measures in multiple areas such as themeasure focus and provided a definition that differentiates the process of medication documentation and medication review from medication reconciliation.

The developer also noted differences in the target population and level of analysis with measure 0553 and the 5 related measures. 0553 addresses the health plan level of analysis for older adults age 65 years and older.

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

5. Related and Competing

Comments

** There are medication reconciliation measures which are not directly competing. They take medication review to the next step.

**None.

** There are I believe with dialysis patients and new Medicaid measures?

** There are a number of competing measures listed; however all occur in different parts of the healthcare continuum. The other measures also focus on low-level documentation functions of medciation reconciliation versus therapeutic medication review.

**No.

** 5 related or competing measures listed.

**Several measures but most are inpatient or hemodialysis.

** There are competing measures that seem more focused on specific conditions - admission or discharge for example. This measure is more global and does overlap these measures.

** Standardized operational definition of "med review" and "med review process" sorely needed.

** Harmonization should be attempted.

** There are related measures but have been harmonizedas much as possible but this is a different measure.

** 0097 : Medication Reconciliation Post-Discharge • 0419e : Documentation of Current Medications in the Medical Record • 2456 : Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient • 2988 : Medication Reconciliation for Patients Receiving Care at Dialysis Facilities • 3317 : Medication Reconciliation on Admission. The developer noted that this measure has been harmonized as much as possible and that measure 0553 s differs from other related measures in multiple areas such as the measure focus and provided a definition that differentiates the process of medication documentation and medication review from medication reconciliation.

**several global measure based on age rather than comorbidity and frailty

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

- **No NQF members who have submitted a support/non-support choice**

Brief Measure Information

NQF #: 0553

Corresponding Measures:

De.2. Measure Title: Care for Older Adults (COA) – Medication Review

Co.1.1. Measure Steward: National Committee for Quality Assurance

De.3. Brief Description of Measure: Percentage of adults 65 years and older who had a medication review during the measurement year. A medication review is a review of all a patient's medications, including prescription medications, over-the-counter (OTC) medications and herbal or supplemental therapies by a prescribing practitioner or clinical pharmacist.

1b.1. Developer Rationale: Medication review can be a useful tool to reduce medication related problems (Christensen & Lundh 2012). The process of reviewing a patient's medication list reduces the risk of adverse drug interactions being overlooked and helps physicians minimize the duplication and complexity of the patient's medication regimen (Kallio et al. 2018). This in turn may increase patient adherence to the medication regimen and reduce hospital readmission rates. A recent systematic review found that medication review resulted in a decrease in the number of drug-related problems among patients (Huiskes et al. 2017).

Christensen, M., & Lundh, A. (2016). Medication review in hospitalised patients to reduce morbidity and mortality. Cochrane Database of Systematic Reviews, (2).

Huiskes, V. J. B., Burger, D. M., van den Ende, C. H. M., & van den Bemt, B. J. F. (2017). Effectiveness of medication review: a systematic review and meta-analysis of randomized controlled trials. BMC family practice, 18(1), 1-15.

Kallio, S. E., Kiiski, A., Airaksinen, M. S., Mäntylä, A. T., Kumpusalo-Vauhkonen, A. E., Järvensivu, T. P., & Pohjanoksa-Mäntylä, M. K. (2018). Community Pharmacists´ Contribution to Medication Reviews for Older Adults: A Systematic Review. Journal of the American Geriatrics Society, 66(8), 1613-1620.

Knight, E.L., J. Avorn. Quality indicators for appropriate medication use in vulnerable elders. Ann. Intern. Med. 2001. 703-10.

S.4. Numerator Statement: At least one medication review conducted by a prescribing practitioner or clinical pharmacist during the measurement year and the presence of a medication list in the medical record.

S.6. Denominator Statement: All patients 66 years and older as of the end (e.g., December 31) of the measurement year.

S.8. Denominator Exclusions: Exclude members who use hospice services.

De.1. Measure Type: Process

S.17. Data Source: Claims, Electronic Health Records, Paper Medical Records

S.20. Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 05, 2009 **Most Recent Endorsement Date:** Aug 10, 2012

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

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

0553_-_nqf_evidence_attachment_7.1.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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.

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0553

Measure Title: Care for Older Adults (COA) – Medication Review

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: Click here to enter a date

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- <u>Outcome</u>: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- <u>Intermediate clinical outcome</u>: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.

- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☐ Outcome: Click here to name the health outcome

☐ Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Medication review of older adults

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Adults age 65 and older >> medication review performed and medication list documented in the medical record >> reduction in adverse drug events >> improved health outcomes

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

N/A – this measure is not derived from patient report

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

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

N/A

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Effectiveness of Medication Review: A Systematic Review and Meta-Analysis of Randomized Controlled Trials** • **Victor Johan Bernard Huiskes, David Marinus Burger, Cornelia Helena Maria van den Ende, Bartholomeus Johannes Fredericus van den Bemt** • **2017** • **Huiskes, V. J. B., Burger, D. M., van den Ende, C. H. M., & van den Bemt, B. J. F. (2017). Effectiveness of medication review: a systematic review and meta-analysis of randomized controlled trials. BMC family practice, 18(1), 1-15.** • **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5240219/**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	Results: "Medication review resulted in a decrease in the number of drug-related problems, more changes in medication, more drugs with dosage decrease and a greater decrease or smaller increase of the number of drugs."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Studies were evaluated according to their risk of bias, using Cochrane criteria. Seventeen out of 31 studies (55%) met the criteria for low risk of bias.

Provide all other grades and definitions from the evidence grading system	This systematic review, assessing the effectiveness of medication review, follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram. The checklist includes items deemed essential for transparent reporting of a systematic review. The Explanation and Elaboration document cited here explains the meaning and rationale for each checklist item. For additional information, see: Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100.
Grade assigned to the **recommendation** with definition of the grade	N/A This systematic review did not include recommendations.
Provide all other grades and definitions from the recommendation grading system	N/A This systematic review did not include recommendations.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• Quantity- 31 randomized clinical trials were included in this systematic review. • Quality – All studies included were randomized clinical trials.
Estimates of benefit and consistency across studies	The findings of this systematic review are in line with the findings of other systematic reviews assessing the effect of medication review.
What harms were identified?	No specific harms were identified.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new randomized controlled trials isolating the effect of medication review in ambulatory settings have been conducted.

1a.4 OTHER SOURCE OF EVIDENCE

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

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

N/A – A Systematic Review is the source of evidence cited in 1a.3.

2012 Submission

Medication review can be a useful tool to reduce medication related problems (Krska 2001). Elderly patients possess several factors, including chronic conditions and increased drug utilization, which makes them particularly prone to adverse drug events resulting from multiple care settings (Marcum 2010).

Hospital medication records for admitted patients are often incomplete. A comparison of medication histories maintained by the hospital for admitted patients with community pharmacy records revealed that the hospital's records omitted 26% of the medications in use. This study also found that 61% of all patients had one or more drugs that were not registered with the hospital (Lau 2000). Significant changes can occur to a patient's medications during hospitalization; a study by Beers et al. found that 45% of all discharge medications were initiated during hospitalization (1989).

The process of resolving discrepancies in a patient's medication list reduces the risk of adverse drug interactions being overlooked and helps physicians minimize the duplication and complexity of the patient's medication regimen (Wenger 2004). This in turn may increase patient adherence to the medication regimen and reduce hospital readmission rates. A study by Gillespie et al utilized a randomized pharmacist-led medication review process of hospitalized patients and demonstrated a subsequent 16% reduction in all visits to the hospital and a 47% reduction in visits to the ED (Gillespie 2009).

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

N/A

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

N/A

2012 Submission

Krska J, Cromarty JA, Arris F, et al. Pharmacist-led medication review in patients over 65: a randomized, controlled trial in primary care. Age Ageing, 2001;30:205-211.
Marcum ZA, Handler SM, Boyce R, et al. Medication Misadventures in the Elderly: A Year in Review. Am J Geriatr Pharmacother. 2010;8:77-83.
Lau HS, Florax C, Porsius AJ, De Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. Br J Clin Pharmacol. 2000;49(6):597-603.
Beers MH, Dang J, Hasegawa J, Tamai IY. Influence of hospitalization on drug therapy in the elderly. J Am Geriatr Soc. 1989;37(8):679-83.
Wenger NS and Young R. Working paper: Quality Indicators of Continuity and Coordination of Care for Vulnerable Elder Persons. Rand: August 2004.
Gillespie U, Alassaad A, Henrohn D, et al. A Comprehensive Pharmacist Intervention to Reduce Morbidity in Patients 80 Years or Older. Arch Intern Med. 2009;169:894-900.

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.

Medication review can be a useful tool to reduce medication related problems (Christensen & Lundh 2012). The process of reviewing a patient's medication list reduces the risk of adverse drug interactions being overlooked and helps physicians minimize the duplication and complexity of the patient's medication regimen (Kallio et al. 2018). This in turn may increase patient adherence to the medication regimen and reduce hospital readmission rates. A recent systematic review found that medication review resulted in a decrease in the number of drug-related problems among patients (Huiskes et al. 2017).

Christensen, M., & Lundh, A. (2016). Medication review in hospitalised patients to reduce morbidity and mortality. Cochrane Database of Systematic Reviews, (2).

Huiskes, V. J. B., Burger, D. M., van den Ende, C. H. M., & van den Bemt, B. J. F. (2017). Effectiveness of medication review: a systematic review and meta-analysis of randomized controlled trials. BMC family practice, 18(1), 1-15.

Kallio, S. E., Kiiski, A., Airaksinen, M. S., Mäntylä, A. T., Kumpusalo-Vauhkonen, A. E., Järvensivu, T. P., & Pohjanoksa-Mäntylä, M. K. (2018). Community Pharmacists´ Contribution to Medication Reviews for Older Adults: A Systematic Review. Journal of the American Geriatrics Society, 66(8), 1613-1620.

Knight, E.L., J. Avorn. Quality indicators for appropriate medication use in vulnerable elders. Ann. Intern. Med. 2001. 703-10.

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 following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data is summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data is stratified by year.

Medicare

Measurement Year: 2016; 2015; 2014

Number of Healh Plans: 504; 427; 404

Mean: 88.0; 84.1; 87.2

STDEV: 16.1; 18.8; 14.0MIN: 0.1; 0.0; 0.0

MAX: 100.0; 100.0; 100.0

P10: 71.3; 57.9; 73.5

P25: 86.5; 80.3; 82.0

P50: 93.2; 91.0; 90.7

P75: 97.5; 96.6; 95.8

P90: 99.5; 98.9; 98.6

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.

N/A

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.*

HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a

plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities.

Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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

While studies have identified disparities in medication prescribing practices (CMS and RAND 2017; Hall-Lipsy 2010) and rates of medication adherence (Lewey 2013), to our knowledge no studies have found disparities in rates of medication review (the focus of this measure) by race/ethnicity, gender, age, insurance status, socioeconomic status and/or disability. As one example, a recent study that examined medication review found no differences by race (Trivedi et al. 2018).

CMS Office of Minority Health and RAND Corporation. Gender Disparities in Health Care in Medicare Advantage. Baltimore, MD. 2017.

Hall-Lipsy EA, Chisholm-Burns MA. Pharmacotherapeutic disparities: racial, ethnic, and sex variations in medication treatment. Am J Health Syst Pharm. 2010 Mar 15;67(6):462-8.

Lewey J, Shrank WH, Bowry AD, Kilabuk E, Brennan TA, Choudhry NK. Gender and racial disparities in adherence to statin therapy: a meta-analysis. Am Heart J. 2013 May;165(5):665-78, 678.e1.

Trivedi M, Fung V, Kharbanda EO, Larkin EK, Butler MG, Horan K, Lieu TA, Wu AC. Racial disparities in family-provider interactions for pediatric asthma care. J Asthma. 2018 Apr;55(4):424-429.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

Care Coordination

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

Elderly, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions

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

N/A

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** 0553_COA_Med_Review_Value_Sets.xlsx

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

No, this is not an instrument-based measure **Attachment:**

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.

Not an instrument-based measure

S.3.1. <u>For maintenance of endorsement:</u> 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. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

No significant changes.

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.*

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

At least one medication review conducted by a prescribing practitioner or clinical pharmacist during the measurement year and the presence of a medication list in the medical record.

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)*

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

This measure can be met using the administrative specification (using administrative claims codes) or the hybrid specification (using administrative claims codes and medical record review).

Administrative: Either of the following meet criteria:

- Both of the following during the same visit during the measurement year where the provider type is a prescribing practitioner or clinical pharmacist:
  - At least one medication review (Medication Review Value Set).
  - The presence of a medication list in the medical record (Medication List Value Set).
- Transitional care management services (Transitional Care Management Services Value Set).

Exclude services provided in an acute inpatient setting (Acute Inpatient Value Set; Acute Inpatient POS Value Set).

(See corresponding Excel document for the value sets referenced above.)

Hybrid: Documentation must come from the same medical record and must include one of the following:

- A medication list in the medical record, and evidence of a medication review by a prescribing practitioner or clinical pharmacist and the date when it was performed.
- Notation that the member is not taking any medication and the date when it was noted.

A review of side effects for a single medication at the time of prescription alone is not sufficient. An outpatient visit is not required to meet criteria. Do not include medication lists or medication reviews performed in an acute inpatient setting.

Prescribing practitioner is defined as a practitioner with prescribing privileges, including nurse practitioners, physician assistants and other non-MDs who have the authority to prescribe medications.

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

All patients 66 years and older as of the end (e.g., December 31) of the measurement year.

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).

Use administrative data to identify all patients 66 years and older as of the end of the measurement year.

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

Exclude members who use hospice services.

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.)*

Exclude members who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These members may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

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.)*

N/A

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

No risk adjustment or risk stratification

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.)*

Step 1. Determine the eligible population: All patients 66 years and older as of the end (e.g., December 31) of the measurement year.

Step 2: Identify the denominator: Exclude any patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

The remainder is the eligible population

Step 3: Identify the numerator: Individuals in the denominator who have documentation of at least one medication review conducted by a prescribing practitioner or clinical pharmacist and have a medication list in their medical record.

Step 4: Calculate the rate: Numerator/Denominator

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.

N/A

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.

N/A – This measure is not based on a survey or instrument.

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

If other, please describe in S.18.

Claims, Electronic Health Records, Paper Medical Records

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.

This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via NCQA's online data submission system.

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)*

No data collection instrument provided

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

Health Plan

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

Outpatient Services

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

nqf_testing_attachment_7.1_COA.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.

No - This measure is not risk-adjusted

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.

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

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

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

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.17*)**	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

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*).

2018 Submission

N/A

1.3. What are the dates of the data used in testing? 01/01/2016 – 12/31/2016

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

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.20*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

This measure assesses whether adults aged 66 and older have had a medication review during the measurement year. A medication review includes a review of all of the member's medications, including prescription medications, over-the-counter medications and herbal or supplemental therapies by a prescribing practitioner or clinical pharmacist. To meet the numerator criteria, the member must have had a medication review and also have a current medication list documented in the medical record. The intended use of the measure is to assess the quality of care in health plans serving older adults.

MEASURE SCORE RELIABILITY TESTING

The measure score reliability was calculated from HEDIS data that included 504 Medicare Advantage Special Needs Plans (SNPs). The measured entities included all Medicare Advantage Special Needs Plans (SNPs) submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

CONSTRUCT VALIDITY TESTING

Construct validity was calculated from HEDIS data that included 504 Medicare Advantage Special Needs Plans (SNPs). The measured entities included all Medicare Advantage Special Needs Plans (SNPs) submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

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*)

2018 Submission

PATIENT SAMPLE FOR MEASURE SCORE RELIABILITY AND FOR VALIDITY TESTING

In 2016, HEDIS measures covered 17.6 million Medicare beneficiaries, of which about 2 million were enrolled in Special Needs Plans. Data are summarized at the health plan level. Below is a description of the sample used for measure score reliability and validity testing. It includes number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans. In 2016, 505

Needs Plans were required to report HEDIS measures, and our sample includes 504 of those plans.

Product Type	Number of Plans	Median number of eligible members per plan
Medicare Special Needs Plan	504	338

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.

2018 Submission

N/A. The same data was used for reliability and validity testing of this measure.

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.

2018 Submission

We did not analyze performance by social risk factors.

2a2. RELIABILITY TESTING

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

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

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

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

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

2018 Submission

Same as previous (2012).

2012 Submission

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

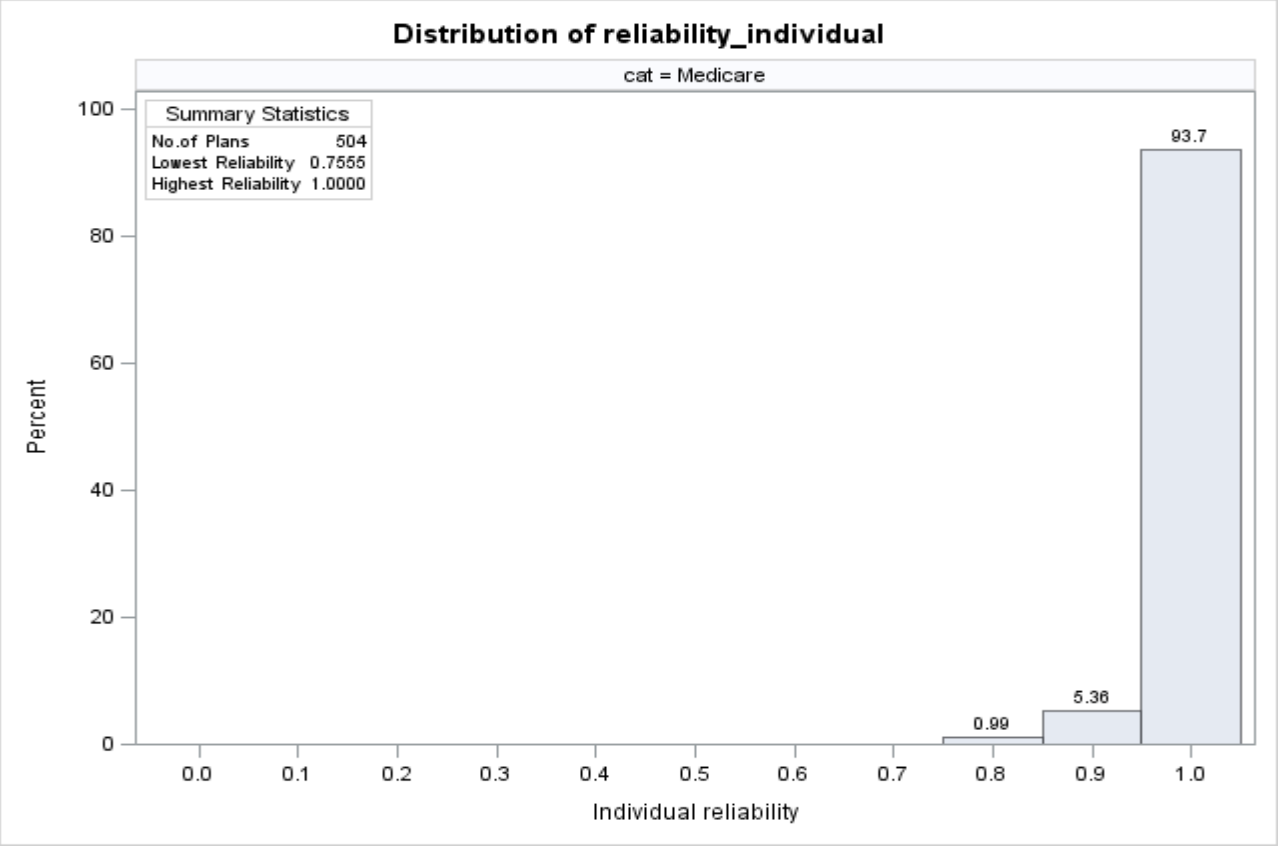
2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

2018 Submission

Reliability for this measure was calculated as 0.985 using 2016 data.

Beta-Binomial Statistic

0.985



2012 Submission

Reliability for this measure was calculated as 0.98712.

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?*)

2018 Submission

Testing suggests the measure has strong reliability. The beta binomial result of 0.985 exceeds the 0.7 threshold.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

2018 Submission

We assessed construct validity and face validity for this measure.

Method of Assessing Construct Validity: We tested for construct validity by exploring whether the Care for Older Adults – Medication Review measure is correlated with the Care for Older Adults – Pain Assessment measure. We hypothesized that organizations that perform well on the Care for Older Adults – Medication Review measures should also perform well on Care for Older Adults – Pain Assessment measure, because the two measures both focus on older adult patients and they both include a key component of assessment for this population. Both measures are reported by the same type of organization (Special Needs Plans), further ensuring a fair comparison.

To test this correlation, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

Method of Assessing Face Validity: We describe below NCQA's process for both measure development and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's board of directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification Support (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assures that measures we use are valid.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs – whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. On average, NCQA receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQA's Board of Directors will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported, and audited before it is used for public accountability or accreditation. This is not testing – the measure was already tested as part of its development – rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated, or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

2012 Submission

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement. This panel included representatives from key stakeholder groups geriatricians, health plans, Medicare officials and researchers. Experts reviewed the results of the field test and assessed whether the results were consistent with expectations, whether the measure represented.

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

2018 Submission

STATISTICAL RESULTS OF CONSTRUCT VALIDITY TESTING

The results in Table 1 indicate that there is a strong, positive relationship between the Care for Older Adults – Medication Review measure and the Care for Older Adults – Pain Assessment measure. This relationship is statistically significant (p<0.0001).

TABLE 1. Pearson Correlation between Care for Older Adults – Medication Review measure and Care for Older Adults – Pain Assessment measure (2016 data)

	Pearson Correlation Coefficient
	Care for Older Adults – Pain Assessment measure
Care of Older Adults – Medication Review measure	0.82

Note: Correlation is significant at p<0.0001

RESULTS OF FACE VALIDITY ASSESSMENT

Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity.

2012 Submission

This measure was deemed valid by the expert panel.

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?*)

2018 Submission

INTERPRETATION OF CONSTRUCT VALIDITY ASSESSMENT

The two measures had high correlation, which indicates the measure has good construct validity.

INTERPRETATION OF FACE VALIDITY ASSESSMENT

Input from NCQA's multi-stakeholder measurement advisory panels and those submitting to public comment agree that *Care for Older Adults – Medication Review* measure is measuring what it intends to measure and the results of measurement allow users to accurately differentiate quality across health plans.

2012 Submission

FACE VALIDITY

Multiple NCQA panels concluded with good agreement that the measures are specified as accurately as possible. This measure meets the test for face validity.

2b2. EXCLUSIONS ANALYSIS

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

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

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*)

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*)

_____

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.

2018 Submission

N/A. Not an intermediate or health outcome, or PRO-PM, or resource use measure.

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

☒ **No risk adjustment or stratification**

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

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

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

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

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?

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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.**

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

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

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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*)

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)*

2018 Submission

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans' performance is significantly different from each other.

2012 Submission

Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

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*)

2018 Submission

HEDIS 2017 VARIATION IN PERFORMANCE ACROSS HEALTH PLANS (Data from 2016 measurement year)

	Avg. EP	Avg. Performance (%)	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicare	338	88.0	16.1	71.3	86.5	93.2	97.5	99.5	11.0	<0.0001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile Range

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

2012 Submission

Medicare

Measurement Year: 2010; 2009; 2008

N: 316; 314; 341

MEAN: 65.4; 60.6; 57.7

STDEV: 22.1; 25.3; 26.6

STDERR: 1.24; 1.43; 1.44

MIN: 0; 0; 0

MAX; 100; 100; 100

P10: 38.8; 19.9; 10.5

P25: 53; 46.5; 49.9

P50: 67.1; 67; 63.6

P75: 81.9; 78.2; 74.7

P90: 93.2; 90.5; 87.4

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?*)

2018 Submission

The IQR shows that there is 11.0 percentage point gap between the 25th and 75th percentile, and that this difference is statistically significant (p<.0001). This gap represents an average of 37 more patients that have had their medications reviewed in high-performing plans compared to low-performing plans (estimated from average health plan eligible population), which is a meaningful difference.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

This measure has only one set of specifications.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

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*)

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*)

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2018 Submission

This measure is collected with a complete sample.

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*)

2018 Submission

This measure is collected with a complete sample.

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

2018 Submission

This measure is collected with a complete sample; there are no missing data on this measure.

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 by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, 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).

To allow for widespread reporting across health plans, this measure is collected through multiple data sources (administrative data, electronic health data, and paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

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

Attachment:

3c. Data Collection Strategy

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

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

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

NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable stakeholders to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:

1) Information practices and control procedures

2) Sampling methods and procedures

3) Data integrity

4) Compliance with HEDIS specifications

5) Analytic file production

6) Reporting and documentation

In addition to the HEDIS audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system informs both annual updates to the measures as well as routine

re-evaluation of measures. These processes include updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence.

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)*.

Broad public use and dissemination of these measures, without modification, are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Modifications to, and/or commercial use of, a measure requires the prior written consent of NCQA and is subject to a license at the discretion of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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 CMS Star Ratings for Medicare Advantage Plans and Prescription Drug Plans https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html Payment Program CMS Star Ratings for Medicare Advantage Plans and Prescription Drug Plans https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html Quality Improvement (external benchmarking to organizations) Healthcare Effectiveness Data and Information Set https://www.ncqa.org/hedis/

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

CMS STAR RATINGS FOR MEDICARE ADVANTAGE PLANS AND PRESCRIPTION DRUG PLANS: This measure is included in the composite Medicare Advantage Star Rating. CMS calculates a Star Rating (1-5) for all Medicare Advantage health plans based on 48 performance measures. Medicare beneficiaries can view the star rating

and individual measure scores on the CMS Plan Compare website. The Star Rating is also used to calculate bonus payments to health plans with excellent performance. The accountable entity is the health plan, and the service assessed in this measure can take place in any health care setting except acute inpatient care. Medicare Advantage covers 33% of all Medicare beneficiaries (19 million members) across 49 states.
The HEALTHCARE EFFECTIVENESS AND DATA INFORMATION SET (HEDIS) is one of health care's most widely used performance measurement and improvement tools. It includes measures for physicians, PPOs, and other organizations. Individual HEDIS measures are used in numerous public reporting, payment, accreditation, and quality improvement programs. 184 million individuals are enrolled in health plans or other organizations that report HEDIS measures (approximately 56% of the total U.S. population).

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?*)
N/A

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.

Special Needs health plans that report HEDIS calculate their rates and know their performance when submitting results. CMS publicly reports rates for these plans to help Medicare Managed Care Organizations understand how they perform relative to others.

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.

CMS publicly reports average performance rates on this measure annually.

(See https://www.cms.gov/Medicare/Health-Plans/SpecialNeedsPlans/Downloads/2016-HEDIS-Report.pdf for a recent example.) NCQA also presents data at various conferences and webinars annually. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. In addition, NCQA regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.

NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

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

Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about who can perform the care that meets the measure numerator.

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

This measure has been deemed a priority measure by CMS, as illustrated by its use in accountability programs such as the Medicare Advantage Star Ratings program.

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.

During the measure's last update, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure to exclude patients on hospice to eliminate undue burden on these patients. We also revised the measure to focus medication reviews on members in ambulatory care. The intent of this measure is to review the medications that an individual regularly takes in community settings (rather than medications prescribed and monitored in acute care).

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.

Over the past three years, this measure has shown slight improvement across health plans (see section 1b.2 for summary of data from health plans). These data are nationally representative and indicate that a higher percentage of individuals enrolled in these health plans are receiving high quality care compared when compared with scores prior to the past 3 years.

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.

There have been no identified unintended findings for this measure during testing or since implementation.

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

There have been no identified unexpected benefits for this measure during testing or since implementation.

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)

0097 : Medication Reconciliation Post-Discharge

0419 : Documentation of Current Medications in the Medical Record

2456 : Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

2988 : Medication Reconciliation for Patients Receiving Care at Dialysis Facilities

3317 : Medication Reconciliation on Admission

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?

Yes

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

See response in 5b.1 (response would not fit in this text box).

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.)

ANSWER TO 5A.1:

NCQA is committed to harmonization across measures and reducing unnecessary burden in measurement. However, it is important to note that the numerator (the specific health care service) being reported in this measure (Measure 0553) differs from many of the other related measures.

Measures 0097, 2456, 3317, and 2988 address MEDICATION RECONCILIATION, which is a care service that includes compiling a list of medications the patient is currently taking and comparing it against a second list (generally a physician's admission, transfer, and/or discharge orders) in order to reconcile discrepancies between the two lists and make sure the patient is prescribed the appropriate medications and to decrease the likelihood of adverse medication interactions.

This care service is different from a MEDICATION REVIEW, which is the focus of this submission (Measure 0553). In a medication review, the goal is a critical examination of all the medications a patient is taking with the objective of reaching an agreement with the patient about treatment, optimizing the impact of medicine, and minimizing medication-related problems.

A medication review is also different from a simple documentation of current medications in the medical record (the focus of Measure 0419e), because this measure involves a review of medications in addition to a documentation of the patient's medications in the medical record.

Additional differences among the measures include level of accountability and target population, as demonstrated below:
0053: Care for Older Adults – Medication Review
Level of accountability: Health plan
Target population: Older adults (age 65 years and older)
0097: Medication Reconciliation Post Discharge
Level of accountability: Health plan
Target population: Adults 18+ discharged from hospital
0419e: Documentation of Current Medications in the Medical Record
Level of accountability: Individual clinician
Target population: Adults 18+
2456: Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient
Level of accountability: Facility (hospital)
Target population: Adults 18+ discharged from hospital
3317: Medication Reconciliation on Admission
Level of accountability: Facility (hospital)
Target population: Adults 18+ admitted to hospital
2988: Medication Reconciliation for Patients Receiving Care at Dialysis Facilities
Level of accountability: Facility (dialysis facility)
Target population: Adults permanently assigned to a dialysis facility
Evidence of performance gap and relation to risk of adverse events:

• Many medication errors occur during times of transition, when patients receive medications from different prescribers who lack access to patients' comprehensive medication list. Conducting medication reconciliation at major care transitions (eg, upon admission, upon discharge) may improve patients' ability to manage their medication regimen properly and reduce the number of medication errors (Measures #0097, 2456, 3317, 2988).
• Older adults are a vulnerable population and are more likely to have multiple comorbid conditions and thus be receiving multiple medications. This places them at higher risk of an adverse medication event, even without a care transition. This supports an annual medication review targeted specifically to older adults (Measure #0053). This measure is more specifically targeted to a vulnerable population and less burdensome to providers than a medication list documented at every medical visit (Measure #0419e).

ANSWER TO 5b.1:
While the other measures generally address a similar focus (medications), no other NQF-endorsed measures address both the same measure focus AND the same target population.

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.

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): National Committee for Quality Assurance

Co.2 Point of Contact: Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Co.3 Measure Developer if different from Measure Steward: National Committee for Quality Assurance

Co.4 Point of Contact: Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.

Geriatric Measurement Advisory Panel

Arlene Bierman,MD, MS, AHRQ

Patricia Bomba,MD, MACP, Excellus BlueCross BlueShield

Jennie Chin Hansen, RN, American Geriatrics Society (Retired)?

Joyce Dubow, MUP,Public Member

Peter Hollmann,MD, Brown University

Steven Phillips,MD, CMD, Geriatric Specialty Care

Wade Aubry,MD, UCSF Institute for Health Policy

Jane Sung, JD, AARP

Eric Tangalos, MD, FACP, AGSF, CMD,Mayo Clinic

Dirk Wales,MD, PsyD, Cigna HealthSpring

Neil Wenger,MD, UCLA

Nicole Brandt, PharmD, BCPP, CGP, FASCP,UMD Pharmacy

Karen Nichols, MD,Amerihealth Caritas

Gustavo Ferrer, MD, Aventura Hospital

Jeff Kelman, MMSc, CMS

Joan Weiss, PhD, RN, CRNP, HHS

Committee on Performance Measurement

Andrew Baskin, MD,Aetna

Helen Darling,MA, Strategic Advisor on Health Benefits & Health Care

Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas

Kate Goodrich,MD, MHS, Centers for Medicare and Medicaid Services

David Grossman, MD, MPH,Kaiser Permanente

Christine Hunter, MD, US Office of Personnel Management

Jeffrey Kelman, MMDc, MD, Centers for Medicare & Medicaid Services

Nancy Lane, PhD, Vanderbilt University Medical Center

Bernadette Loftus, MD, The Permanente Medical Group

Adrienne Mims, MD, MPH, Alliant Quality

Amanda Parsons, MD, MBA,Montefiore Health System

Wayne Rawlins, MD, MBA, ConnectiCare

Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic

Eric Schneider, MD, MSc, FACP, The Commonwealth Fund

Marcus Thygeson, MD, MPH, Blue Shield of California

JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

Lina Walker, PhD, AARP Public Policy Institute

Technical Measurement Advisory Panel

Andy Amster, MSPH, Kaiser Permanente

Jennifer Brudnicki, MBA, Geisinger Health Plan

Lindsay Cogan, PhD, MS, New York State Department of Health

Kathy Coltin, MPH, Independent Consultant

Mike Farina, MVP Healthcare

Marissa Finn, MBA, CIGNA HealthCare

Scott Fox, MS, Med, Independence Blue Cross

Carlos Hernandez,CenCal Health

Harmon Jordan,ScD, Westat

Virginia Raney, LCSW, Center for Medicaid and CHIP Services

Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC

Laurie Spoll, Aetna

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2010

Ad.3 Month and Year of most recent revision: 07, 2018

Ad.4 What is your frequency for review/update of this measure? This measure is reviewed approximately every 3 years, and sooner if clinical guidelines or evidence

Ad.5 When is the next scheduled review/update for this measure? 12, 2019

Calculated measure results, based on unadjusted HEDIS specifications, may not be termed “Health Plan HEDIS rates” until they are audited and designated reportable by an NCQA-Certified Auditor. Such unaudited results should be referred to as “Unaudited Health Plan HEDIS Rates.” Accordingly, “Heath Plan HEDIS rate” refers to and assumes a result from an unadjusted HEDIS specification that has been audited by an NCQA-Certified HEDIS Auditor.

Limited proprietary coding is contained in the measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. NCQA disclaims all liability for use or accuracy of any coding contained in the specifications.

Content reproduced with permission from HEDIS, Volume 2: Technical Specifications for Health Plans. To purchase copies of this publication, including the full measures and specifications, contact NCQA Customer Support at 888-275-7585 or visit www.ncqa.org/publications.

Ad.7 Disclaimers: This HEDIS® performance measure is not a clinical guideline and does not establish a standard of medical care and has not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 Additional Information/Comments: NCQA Notice of Use. Broad public use and dissemination of these measures, without modification, are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Modifications to, and/or commercial use of, a measure requires the prior written consent of NCQA and is subject to a license at the discretion of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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

Brief Measure Information

NQF #: 0555

Measure Title: INR Monitoring for Individuals on Warfarin

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: Percentage of individuals at least 18 years of age as of the end of the measurement period with at least 56 days of warfarin therapy who receive at least one International Normalized Ratio (INR) test during each 56-day interval with active warfarin therapy.

Developer Rationale: Warfarin remains the most commonly prescribed anticoagulant in the United States overall[1] and among Medicare PDP beneficiaries.[2] Warfarin has a narrow therapeutic range requiring regular monitoring with the INR test and dose adjustment to maintain patient safety by avoiding thromboembolism or bleeding complications. Warfarin has been identified as the leading drug class implicated in emergency hospitalizations for adverse drug events in adults over 65 years of age.[3] Consequences of adverse drug events related to warfarin therapy are serious and can be fatal. One study found a case-fatality rate of 11.3% for venous thromboembolism (VTE).[4] Case fatality rates for patients with major bleeding can range from 8 percent to 11 percent[4-7] and can reach 45 percent to 50 percent for those with intracranial bleeding.[8,9] For patients with stable INRs, clinical practice guidelines recommend frequent and continuous INR monitoring every 4 to 12 weeks.[10,11] This measure aims to promote patient safety through medication management of individuals on warfarin and to encourage providers to conduct regular INR monitoring for these individuals. Regular INR monitoring is associated with increased time in therapeutic range [12-14] and reduced risk of thromboembolism,[14] whereas subtherapeutic INR is correlated with significantly higher total healthcare costs[15, 16] and greater risks of stroke/SE,[17] major bleeding[17,18], thromboembolism,[18] and mortality.[17-19]

Current health plan-level performance indicates a quality gap remains. Using 2016 QHP claims data, we found there is a 15.2% difference between the 10th and 90th percentiles with a median score of 56.6% indicating that just over half of health plan members receive regular INR monitoring. In 2016 Medicare claims data, there is an 18.2% difference between the 10th and 90th percentiles with a median score of 71.4% among prescription drug plans. This is a decrease in performance over time compared to the measure developer's previous testing information using data from Medicare prescription drug plans from 2012, which showed a median score of 75.6% and percentiles (P) of performance as follows: P10=64.8%, P25=68.5%, P50=75.6%, P75=81.0%, P90=83.6% indicating variation in performance and room for improvement.[20]

Studies from the literature also suggest an opportunity for improvement in the management of patients on warfarin. A 2015 retrospective study of 9,433 patients who received warfarin for >6 months found that 39% of INR values were out of range.[15] A 2016 review of 6 meta-analyses evaluating the stability of INR (i.e., greater than or equal to 65% time in therapeutic range [TTR]) for patients on anticoagulation therapy found that there is high variability among patients and when patients achieve the target INR range, they do not remain stable

and typically have INR values below the therapeutic range, increasing their risk of adverse drug events.[21] A study published in 2018 provides support for the process-outcome linkage: "Patients with TTR <65% had a higher risk for any stroke/SE (HR: 1.57; 95% CI: 1.41–1.75), major bleeding (HR: 2.78; 95% CI: 2.55–3.03) and all-cause mortality (HR: 1.73; 95% CI: 1.67–1.79)."[17] These findings are similar to another study that found that INR variability was shown to be a predictor of mortality where patients with more TTR had higher survival time.[19] The association between TTR and thromboembolism, major bleeding, and death has also been demonstrated in a sample of patients with mechanical heart valve prosthesis.[18]

The literature combined with our empirical evidence suggests room for improvement in anticoagulation management which this measure supports through INR monitoring by specifying an evidence-based interval of 56 days (8 weeks).[12] Further, NQF 0555 is the only endorsed measure that addresses regular monitoring for individuals on warfarin. While NQF 0555 is related to both NQF 0556 (INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications) and NQF 2732 (INR Monitoring for Individuals on Warfarin after Hospital Discharge), all three measures have different clinical foci and target populations. These measures are discussed further in question 5a.2

Numerator Statement: The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy. The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.

Denominator Statement: Continuously enrolled individuals, at least 18 years of age at of the end of the measurement period, with at least 56 days of warfarin therapy during the measurement period.

Denominator Exclusions: 1. Individuals who are monitoring INR at home. These individuals are excluded because the claims associated with home INR monitoring are associated with up to four INR tests per claim. Therefore, a single claim for home INR monitoring would not be representative of a single INR test and would prohibit being able to distinguish if the home INR test was within the 56-day timeframe specified by the numerator of this measure.

2. Individuals who have first or last warfarin claims with missing days' supply.

Measure Type: Process

Data Source: Claims

Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 05, 2009 **Most Recent Endorsement Date:** Nov 10, 2014

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 *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary or **Summary of prior review in [year]**

- The developer provided a summary of the link between regular INR monitoring of patients on warfarin and maintaining the patient within the therapeutic range of warfarin with fewer bleeding and thromboembolic events and lower hospitalization and mortality rates.
- The developer provided the following clinical guidelines
    - Management of patients with atrial fibrillation (Compilation of 2006 ACCF/AHA/ESC and 2011 ACCF/AHA/HRS recommendations): A report of the American College of Cardiology/American Heart Associations Task Force on Practice Guidelines
        - 5. INR should be determined at least weekly during initiation of therapy and monthly when anticoagulation is stable."
        - Class I = Benefit >>> Risk. Procedure/Treatment SHOULD be performed/administered.

          Evidence Level: A - Multiple populations evaluated. Data derived from multiple randomized clinical trials or meta-analyses.
        - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the guideline.

    - Evidence-based management of anticoagulant therapy: Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines
        - 3.1. For patients taking VKA therapy with consistently stable INRs, we suggest an INR testing frequency of up to 12 weeks rather than every 4 weeks"
        - Grade 2B: Weak recommendation, moderate-quality evidence
        - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the guideline.
    - 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society
        - 6. Among patients treated with warfarin, the INR should be determined at least weekly during initiation of antithrombotic therapy and at least monthly when anticoagulation (INR in range) is stable.
        - Level of Evidence A: Data derived from multiple randomized clinical trials or meta-analyses.

          Class I: Conditions for which there is evidence and/or general agreement that a given procedure or treatment is useful and effective.
        - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the guideline.

    - The developer provided the following systematic reviews: Anticoagulation intensity and outcomes among patients prescribed oral anticoagulant therapy: A systematic review and meta-analysis.
        - The systematic review concluded, "The risks of hemorrhage and thromboemboli are minimized at international normalized ratios of 2–3. Ratios that are moderately higher than this therapeutic range appear safe and more effective than subtherapeutic ratios."
        - There was no grade assigned for the quality of quoted evidence.
- The developer cited the following evidence to support an INR monitoring interval of up to 12 weeks to support the recommendations and does not change the concussions of the systematic review by Holbrook et al. (2012). Further, Witt et al. (2016) shows the ambiguity in the appropriate length for follow-up.
    - "During the first 3 months of warfarin therapy for VTE we suggest that INR recall intervals not exceed 6 weeks."
    - "For patients demonstrating consistently stable INRs after 3 months of warfarin therapy for VTE we suggest that INR recall intervals can be extended up to 12 weeks."
- There were no unintended harms described in the evidence for INR monitoring.

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:

Exception to evidence

NA

Questions for the Committee:

- The evidence provided by the developer is updated, directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?
- Does the evidence continue to support the 56 day (8-week) time interval for INR Monitoring?
- For structure, process, and intermediate outcome measures:
    - What is the relationship of this measure to patient outcomes?
    - How strong is the evidence for this relationship?
    - Is the evidence directly applicable to the process of care being measured?

Guidance from the Evidence Algorithm

Process measure with systematic review (Box 3) → Summary of the QQC provided (Box 4) → Systematic review concludes moderate quality evidence.

The highest possible rating is "High" for Evidence

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

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.

- Several sources of data were used in testing the measure. Data representing the target population—members enrolled in Affordable Care Act (ACA) Health Insurance Exchange QHP products —are from four issuers, representing seven QHP products in 2015 and eight products in 2016. Patient-level data representing the target population—members enrolled in Affordable Care Act (ACA) Health Insurance Exchange QHP products—were provided to the Measure Developer from one issuer, henceforth Issuer 1. These data were used to calculate all analyses. A data analytic firm provided QHP analytic results for three issuers, henceforth Issuer 2, Issuer 3, and Issuer 4, in lieu of patient-level data.
    - Overall, across 4 QHP products from 3 QHP Issuers with sufficient denominators to report measure rates, the performance scores ranged from 48.9% to 62.1% in 2015, and from 43.9% to 59.1% in 2016.
- Additionally, national claims data from Medicare Part B and stand-alone Part D prescription drug plans (PDPs) were used to supplement the QHP analyses since limited QHP data were available for testing.
    - In 2012, average performance rate of 74.5%. In 2015, average performance rate of 76.7%. In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 71.7%) among Medicare PDPs.
- The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries who take warfarin therapy.

Disparities

- Per developer, among three issuers' QHP products, disparities for sex were not found in either 2015 or 2016 data. In 2015, in one issuer, and in one product, a disparity by age group was evidenced: the 27 to 44 age group had lower performance compared to the reference group of 45 to 64.
- For the Medicare PDP data, although statistical significance was found, national measure rates suggest there is not disparity in care between sexes due to a less than 10% relative difference in measure rates in both 2015 and 2016. However, national measure rates among Medicare PDPs suggest that beneficiaries who were younger, did not identify as white, and were dually eligible for Medicare and Medicaid services had lower measure rates.
- In addition, the developer cited literature that addresses disparities in care.
    - Rose et al. (2013) found that 45% of the 56,490 Veterans Health Administration patients included in their study, who were aged 65 years and older, had at least one gap >=56 days in INR monitoring, representing 44,430 total gaps and 4,482,100 days without INR monitoring over the two-year study period. Predictors of any gaps in monitoring during warfarin therapy that were identified in the study included: younger age (age of 65-69 years versus >=75 years , non-white race (non-Hispanic black race, Hispanic race, and Native American race, and residence in a zip code with a poverty level below the federal poverty line (poverty level 17.8%-100.0%).
    - Witt et al. (2013) study found that factors associated with nonadherence to INR testing included: younger age, and male sex.

Questions for the Committee:

- Specific questions on information provided for gap in care.
- 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
Comments:

**Updated evidence was provided concerning the importance of this measure to patient safety.

**A process metric, logic: INR monitoring lowers thromboembolic events and hospitalizations, evidence level A.

**Yes.

**The evidence is directly related to the process of concern. New evidence does support the potential to increase the 12-week time interval for INR monitoring be considered.

**Systemic reviews consistent, graded. Guidelines Class 1, Evidence level A.

**Agree.

**Moderate level of evidence.

**Acceptable.

**Moderate - evidence exists to support 84 day interval for some stable patients.

**Moderate evidence provided and no need for futher discussion of evidence.

**Percentage of individuals at least 18 years of age as of the end of the measurement period with at least 56 days of warfarin therapy who receive at least one International Normalized Ratio (INR) test during each 56-day interval with active warfarin therapy. The developer submitted a systematice evidence review; quality, consistency and quantity of the evidence; graded the evidence. No need for repeat discussion and vote on evidence.

1b. Performance Gap

Comments:

**Current performance gap data was provided. A gap is demonstrated and there is room for improvement. Some disparities data was provided.

**High, about 50% compliant and gap between 10th and 90th percentile, no gender disparities but age disparities found.

**Yes.

**A performance gap still exists that warrants a National Quality measure. Disparities were found to exist in younger patients, non-white and dual eligibles.

**Performance rates in the 70s for the medicare population and lower in the ACA population.

**Yes gap and some disparities.

**Performance gap exists. 49-77%.

**Appears still about a toss of a coin that you'll get tested.

**High.

**Ample opportunity for improvement and some data suggestive of disparities.

**Current health plan-level performance indicates a quality gap remains. Using 2016 QHP claims data, the developers found there is a 15.2% difference between the 10th and 90th percentiles with a median score of 56.6% indicating that just over half of health plan members receive regular INR monitoring. In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 71.7%) among Medicare PDPs. The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries who take warfarin therapy. Per developer, among three issuers' QHP products, disparities for sex were not found in either 2015 or 2016 data. In 2015, in one issuer, and in one product, a disparity by age group was evidenced: the 27 to 44 age group had lower performance compared to the reference group of 45 to 64.

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.

Composite measures only:

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

N/A

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

Evaluators: Patient Safety project team staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Link A (Project Team staff)

Developer did updated reliability testing in the maintenance using again approach proposed by Adams (2009) and Scholle et al. (2008) (ratio of signal to noise). Among the QHP products with at least 30 denominator members tested, reliability ranged from 0.60 to 0.79 with a mean reliability of 0.70. For Medicare PDPs, using the method of minimum denominator and volume categories, a minimum of 100 members in the denominator results in an overall reliability score of 0.7. Both results indicate sufficient signal relative to noise to discriminate performance between plans or units of analysis.

For empirical validity testing, the developer did convergent validity testing with Pearson's correlation coefficients and compared the performance of NQF 0555 with NQF 0541 (Proportion of Days Covered [PDC]: 3 Rates by Therapeutic Category). The developer also looked at face validity of NQF 0555. The results indicated the measure is valid. For empirical validity, the performance comparison with NQF 0541 and NQF 0555 were positively correlated at the PDP level. diabetes: r=0.591, hypertension: r=0.700, cholesterol: r=0.751). According to Cohen's thresholds for product-moment correlations, 0.50 or higher is considered a large correlation. For face validity, all responding TEP member (9) agreed that NQF 0555 was valid as specified. 3 TEP members did not complete survey.

Questions for the Committee regarding reliability:

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

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

Evaluation A: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0555

Measure Title: INR Monitoring for Individuals on Warfarin

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.**

No major concerns with the measure specifications or calculation algorithm. However, one question would by why the measure has an optional denominator exclusion of individuals who are in long-term care (LTC) during

the measurement period. Developer responded that this optional exclusion was removed from the current specifications.

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**

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

   Submission document: Testing attachment, section 2a2.2

Developer did testing using 2015–2016 administrative claims data from four issuers (referred to as QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4), containing a total of seven Health Insurance Exchange Qualified Health Plan (QHP) products in 2015 and eight in 2016 and also 2015–2016 administrative claims data from Medicare Parts A, B, and D for beneficiaries enrolled in stand-alone Part D Prescription Drug Plans (referred to as Medicare PDPs). The data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4 included all members with claims associated with the QHP products. QHP products with 500 or fewer total members excluded from all analyses and denominators had to have at least 30 members in order to show the results of the analyses. Note that QHP Issuer 4 did not have sufficient denominator sizes for analyses and is thus not presented in the results section for reliability. The Medicare sample included all beneficiaries from the national Medicare claims database who had at least one month of Part A and Part B coverage and no HMO coverage during the year and who were in a stand-alone Medicare PDP.

Developer did updated reliability testing in the maintenance using again approach proposed by Adams (2009) and Scholle et al. (2008) (ratio of signal to noise). One change to methods used was the exception that they used the method of minimum denominator and volume categories from Scholle et al. instead of the mean denominator.[2] Per developer, this method assumes that the denominator size in each volume category is equal to the minimum for that category and provides a more conservative estimate of reliability for each volume category

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

   Submission document: Testing attachment, section 2a2.3

Among the QHP products with at least 30 denominator members tested (QHP Issuer 1, QHP Issuer 2, QHP Issuer 3), reliability ranged from 0.60 to 0.79 with a mean reliability of 0.70. QHP Issuer 3 had a reliability score less than 0.7. QHP Issuer 3 (Product A) had a reliability score of 0.69 and QHP Issuer 3 (Product B) had a reliability score of 0.60.

For Medicare PDPs, using the method of minimum denominator and volume categories, a minimum of 100 members in the denominator results in an overall reliability score of 0.7.

Both results indicate sufficient signal relative to noise to discriminate performance between plans or units of analysis.

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.**

Reliability methodology and results seem appropriate at the health plan level.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

One question would by why the measure has an optional denominator exclusion of individuals who are in long-term care (LTC) during the measurement period. Per developer, this is because most health plans do not provide coverage for long-term care. This exclusion was not tested due to the lack of coverage for long-term care in the samples.

The exclusion of individuals with home INR monitoring was tested using both Medicare and QHP data. The exclusion is appropriate as individuals monitoring INR at home would not have reliable claims data for INR tests that could be used to satisfy the measure specifications.

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

    Submission document: Testing attachment, section 2b4.

QHP products had a mean measure performance rate of 54.0%, which indicates there is still a quality gap. In addition, performance rates decreased among the Medicare PDPs from 2012 to 2016.

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.

N/A

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

Submission document: Testing attachment, section 2b6.

No concerns. "Days' supply of medication" was complete in the dataset used for testing.

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

This is a process measure so not applicable.

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

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

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

N/A

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**

18. **Method of establishing validity of the measure score:**

☒ **Face validity**

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

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

19. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

The developer did convergent validity testing with Pearson's correlation coefficients and compared the performance of NQF 0555 with NQF 0541 (Proportion of Days Covered [PDC]: 3 Rates by Therapeutic Category). This is an appropriate method for empirical validity of the measure.

The developer also looked at face validity of NQF 0555. TEP members were specifically asked whether they agree with the following statement: "The performance scores resulting from the measure NQF 0555 INR

Monitoring for Individuals on Warfarin, as specified, can be used to distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy."

20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

The results indicated the measure is valid. For empirical validity, the performance comparison with NQF 0541 and NQF 0555 were positively correlated at the PDP level. diabetes: r=0.591, hypertension: r=0.700, cholesterol: r=0.751). According to Cohen's thresholds for product-moment correlations, 0.50 or higher is considered a large correlation.[1]

For face validity, all responding TEP member (9) agreed that NQF 0555 was valid as specified. 3 TEP members did not complete survey.

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

    ☐ **Yes**

    ☐ **No**

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

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

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

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

    ☐ **Low** (NOTE: Should rate LOW if you believe that there 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.)

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

Results for the empirical validity of the measure indicated sufficient results of correlation of the two NQF measures.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

25. **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**

N/A

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

N/A

ADDITIONAL RECOMMENDATIONS

27. **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.**

No additional concerns.

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

2a1. Reliability – Specifications
Comments:

**No concerns.

**Detailed specifications provided.

**Okay.

**Reliability data is sufficient.

**Signal to noise 0.6 to 0.79.

**Claims based reliability seems fine.

**No comments.

**Acceptable.

**No.

**Measure specifications are adequate.

**Developer did updated reliability testing in the maintenance using again approach proposed by Adams (2009) and Scholle et al. (2008) (ratio of signal to noise). Among the QHP products with at least 30 denominator members tested, reliability ranged from 0.60 to 0.79 with a mean reliability of 0.70. This level just meets the NQF requirement for reliability score.

2a2. Reliability – Testing
Comments:

**None.

**Moderate/low, mean score of 0.7, barely meeting threshold.

**No.

**Moderate reliability. No major concerns.

**No.

**No.

**No.

**No.

**Moderate: reliability is on the borderline for acceptability for group comparisons.

**Testing is sufficient.

**The data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4 included all members with claims associated with the QHP products. QHP products with 500 or fewer total members excluded from all analyses and denominators had to have at least 30 members in order to show the results of the analyses. Note that QHP Issuer 4 did not have sufficient denominator sizes for analyses and is thus not presented in the results section for reliability.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:

**None.

**Moderate, face validity and empirical validity.

**No.

**No concerns on validity.

**Pearson's correlation > 0.5.

**No

**No

**No

**Moderate.

**Validity testing is sufficient.

**Measure has face and empirical validity testing. The results indicated the measure is valid. For empirical validity, the performance comparison with NQF 0541 and NQF 0555 were positively correlated at the PDP level. diabetes: r=0.591, hypertension: r=0.700, cholesterol: r=0.751). According to Cohen's thresholds for product-moment correlations, 0.50 or higher is considered a large correlation. For face validity, all responding TEP member (9) agreed that NQF 0555 was valid as specified. No concerns.

**No concerns.

**No concerns, administrative data from claims.

**No

**No concern regarding threats to validity.

**The evidence supports testing Q 6 weeks for the first 3 months then Q 12 weeks. This measure states Q 8 weeks which falls between. The measure states testing once every 8 weeks but does not define whether this is early or late in the treatment.

**No

**No

**Would like to see what high performance on this measure correlates with, eg reduced hospitalizations, head bleeds etc.

**Ok.

**No.

**No threats to the validity of this measure were identified using a limited analysis designed to address missing data. NA re comparability.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment

Comments:

**None.

**No.

**Removal of optional exclusion for LTC is justified and explained.

**Why exclude long term care?

**No risk adjustment-process measure.

**Appears appropriate.

**Exclusion of home monitored patients is appropriate.

**Yes.

**No threats. This measure is not risk-adjusted.

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.

Data Specifications and Elements

- The measure is constructed using administrative claims
- All data elements are in defined fields in electronic claims
- This measure is not an eMeasure. At this time, there is no plan to specify the measure as an eMeasure.

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?
- Is the data collection strategy ready to be put into operational use?

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**High, administrative claims and medical records.

**Do not see a link to results of the testing.

**High feasibility.

**Routinely generated during care delivery. Administrative claims and defined fields in electronic claims.

**Adminstrative data.

**Agree with high feasibility prelim rating.

**No concerns.

**Discuss why social risk factors not conducted.

**No concerns.

**High feasibility. The measure is constructed using administrative claims. All data elements are in defined fields in electronic claims.This measure is not an eMeasure. At this time, there is no plan to specify the measure as an eMeasure.

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

This measure was previously in use for the Quality and Resource Use Reports, but has not been in use since the last NQF review in 2013. Per developer, this measure is now being considered for use in the Quality Rating System for QHPs. The Quality Rating System is intended to inform consumers when choosing a QHP from the Health Insurance Exchange by providing comparisons of the quality of care provided by each health plan. The Quality Rating System is not used for payment or penalty to the health plans. The developer has no further updates at this time on implementation of the measure as the Call Letter for the Quality Rating System and WHP Enrollee Survey has not been announced yet.

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 measure is not currently implemented in a public reporting program, and therefore there is no information available regarding feedback during implementation.

Additional Feedback:

The developer states the a Technical Expert Panel has reviewed the updated measure evidence, testing and performance results. The TEP is comprised of three representatives from large QHP issuers, and nine individuals from other stakeholder groups, such as organization representatives, clinical and nonclinical

experts, and patient/caregiver representatives. Meetings with the TEP were held throughout 2015-2017. TEP members were encouraged to provide feedback throughout the measure re-evaluation process by means of meeting discussions and voting and through follow-up communications. Members were sent a questionnaire focused on face validity and usability which contained closed-ended response options and free text comment fields.

***Questions for the Committee**:*

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

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

RATIONALE: This is a maintenance measure and has not been in an accountability program since 2014. However, the developer states a plan for it to be considered for use in the Quality Rating System for QHPs; there is no specified timeframe provided.

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 measure is not currently implemented in a public reporting program. The developer had provided performance rate of the QHP products and amongst Medicare PDPs which both indicate substantial opportunity for improvement.

QHP products average rate (2016)-54.0%

Medicare PDP average rate (2012)-74.5%

Medicare PDP average rate (2016)-71.7%

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

No unexpected findings provided by developer. This measure is not currently in use, however, previous measure maintenance efforts reported that no unintended negative consequences had been identified in the 2011 Quality and Resource Use Reports.

Potential harms

There are no harms identified by the developer.

Additional Feedback:

- During the last maintenance review of this measure (2015 Patient Safety Final Report), there was discussion and comments by Standing Committee and Public Comments about the 56 day time interval for INR monitoring. The Committee was satisfied with the Developer's rationale/evidence for this 56 day time interval and exclusion rationale for the home health monitoring patients.

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

RATIONALE: The measure is not currently implemented in a public reporting program, however per developer it is under consideration.

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

4a1. Use - Accountability and Transparency
Comments:
**There is no plan outlined for accountability applications, though a plan for including this measure as part of the public's opportunity to evaluate QHPs appears to be under consideration.
**Moderate/low, not reported publically, considered for use in Quality Rating System for QHPs.
**Can you link to results of the testing?
**The measure has not been used since 2014 in an public accountability program. It is in consideration for use as a Quality Rating System measure for QHPs.
**Not currently in use. No feedback.
**Not publically reported.
**Agree with prelim rating.
**Appears to fail NQF expections; this may be a measure that does not really matter, even if it is still a decent enough measure for internal quality improvement.
**OK.
**Not currently publicly reported but planned and feedback was solicited by TEP.
**This measure was previously in use for the Quality and Resource Use Reports, but has not been in use since the last NQF review in 2013. Per developer, this measure is now being considered for use in the Quality Rating System for QHPs. The Quality Rating System is intended to inform consumers when choosing a QHP from the Health Insurance Exchange by providing comparisons of the quality of care provided by each health plan. The Quality Rating System is not used for payment or penalty to the health plans. The developer has no further updates at this time on implementation of the measure as the Call Letter for the Quality Rating System and WHP Enrollee Survey has not been announced yet. This is a maintenance measure and has not been in an accountability program since 2014. However, the developer states a plan for it to be considered for use in the Quality Rating System for QHPs; there is no specified timeframe provided.

4b1. Usability – Improvement
Comments:
** Given the patient safety aspect of this measure, it appears that the measure results should be publicly available for use in choosing a QHP. Benefits outweigh the harm; no unintended consequences.
**No harms identified.
**No but cost should be considered as well as genetic testing?
**Currently, the measure is not in use. Use of warfarin is a high-risk medication concern needing effective monitoring. No known unintended consequences are known for utilization of this measure.
**Considering for use in quality rating system for QHPs on Health Exchange.
**Minimal harm-support for 8 week interval appears sound.
**No harms identified by developer.
**It isn't being used in reality.
**Low: continued debate about 56 day time interval, could penalize providers of some appropriately treated patients.

**no unintended consequences and data presented indicates ample room for improvement however not curently reported so hard to assess.

**Low: The measure is not currently implemented in a public reporting program.

Criterion 5: Related and Competing Measures

Related or competing measures

Related measures:

0556 : INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications-Steward is Centers for Medicare & Medicaid Services

2732 : INR Monitoring for Individuals on Warfarin after Hospital Discharge- Steward is Centers for Medicare & Medicaid Services

Harmonization

All three measures 0555, 0556, and 2732 have the same measure focus, which is INR testing, and their specifications for INR testing are harmonized; however, the three measures have different clinical foci and target populations. Due to the difference in the clinical foci, the timeframe for INR monitoring (three to seven days, 14 days, 56 days) is different among the three measures and complimentary rather than competing with one another.

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

5. Related and Competing

Comments

**No concerns.

**None.

**No.

**Other measures evaluating INR monitoring do not focus on same long-term use of Warfarin, but rather early titration and monitoring activities.

**Harmonized. no competing measures.

**Some overlap but targets a little different-may be able to harmonize in future.

**0556 and 2732 with differing timeframes.

**No concerns.

**There are 2 competing measures with different target populations.

**Related measures but the developer is clear that the other measures have different clinical foci.

**0556 : INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications-Steward is Centers for Medicare & Medicaid Services; 2732 : INR Monitoring for Individuals on Warfarin after Hospital Discharge- Steward is Centers for Medicare & Medicaid Services Harmonization All three measures 0555, 0556, and 2732 have the same measure focus, which is INR testing, and their specifications for INR testing are harmonized; however, the three measures have different clinical foci and target populations. Due to the difference in the clinical foci, the timeframe for INR monitoring (three to seven days, 14 days, 56 days) is different among the three measures and complimentary rather than competing with one another. All three

have the same measure focus, which is INR testing, and their specifications for INR testing are harmonized; however, the three measures have different clinical foci and target populations.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

- No NQF Members have submitted support/non-support choices as of this date.

Brief Measure Information

NQF #: 0555

Corresponding Measures:

De.2. Measure Title: INR Monitoring for Individuals on Warfarin

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

De.3. Brief Description of Measure: Percentage of individuals at least 18 years of age as of the end of the measurement period with at least 56 days of warfarin therapy who receive at least one International Normalized Ratio (INR) test during each 56-day interval with active warfarin therapy.

1b.1. Developer Rationale: Warfarin remains the most commonly prescribed anticoagulant in the United States overall[1] and among Medicare PDP beneficiaries.[2] Warfarin has a narrow therapeutic range requiring regular monitoring with the INR test and dose adjustment to maintain patient safety by avoiding thromboembolism or bleeding complications. Warfarin has been identified as the leading drug class implicated in emergency hospitalizations for adverse drug events in adults over 65 years of age.[3] Consequences of adverse drug events related to warfarin therapy are serious and can be fatal. One study found a case-fatality rate of 11.3% for venous thromboembolism (VTE).[4] Case fatality rates for patients with major bleeding can range from 8 percent to 11 percent[4-7] and can reach 45 percent to 50 percent for those with intracranial bleeding.[8,9] For patients with stable INRs, clinical practice guidelines recommend frequent and continuous INR monitoring every 4 to 12 weeks.[10,11] This measure aims to promote patient safety through medication management of individuals on warfarin and to encourage providers to conduct regular INR monitoring for these individuals. Regular INR monitoring is associated with increased time in therapeutic range [12-14] and reduced risk of thromboembolism,[14] whereas subtherapeutic INR is correlated with significantly higher total healthcare costs[15, 16] and greater risks of stroke/SE,[17] major bleeding[17,18], thromboembolism,[18] and mortality.[17-19]

Current health plan-level performance indicates a quality gap remains. Using 2016 QHP claims data, we found there is a 15.2% difference between the 10th and 90th percentiles with a median score of 56.6% indicating that just over half of health plan members receive regular INR monitoring. In 2016 Medicare claims data, there is an 18.2% difference between the 10th and 90th percentiles with a median score of 71.4% among prescription drug plans. This is a decrease in performance over time compared to the measure developer's previous testing information using data from Medicare prescription drug plans from 2012, which showed a median score of 75.6% and percentiles (P) of performance as follows: P10=64.8%, P25=68.5%, P50=75.6%, P75=81.0%, P90=83.6% indicating variation in performance and room for improvement.[20]

Studies from the literature also suggest an opportunity for improvement in the management of patients on warfarin. A 2015 retrospective study of 9,433 patients who received warfarin for >6 months found that 39% of INR values were out of range.[15] A 2016 review of 6 meta-analyses evaluating the stability of INR (i.e., greater than or equal to 65% time in therapeutic range [TTR]) for patients on anticoagulation therapy found that there is high variability among patients and when patients achieve the target INR range, they do not remain stable and typically have INR values below the therapeutic range, increasing their risk of adverse drug events.[21] A study published in 2018 provides support for the process-outcome linkage: "Patients with TTR <65% had a higher risk for any stroke/SE (HR: 1.57; 95% CI: 1.41–1.75), major bleeding (HR: 2.78; 95% CI: 2.55–3.03) and all-cause mortality (HR: 1.73; 95% CI: 1.67–1.79)."[17] These findings are similar to another study that found that INR variability was shown to be a predictor of mortality where patients with more TTR had higher survival time.[19] The association between TTR and thromboembolism, major bleeding, and death has also been demonstrated in a sample of patients with mechanical heart valve prosthesis.[18]

The literature combined with our empirical evidence suggests room for improvement in anticoagulation management which this measure supports through INR monitoring by specifying an evidence-based interval of 56 days (8 weeks).[12] Further, NQF 0555 is the only endorsed measure that addresses regular monitoring for individuals on warfarin. While NQF 0555 is related to both NQF 0556 (INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications) and NQF 2732 (INR Monitoring for Individuals on Warfarin after Hospital Discharge), all three measures have different clinical foci and target populations. These measures are discussed further in question 5a.2

Citations

1. US Department of Health and Human Services, Office of Disease Prevention and Health Promotion. National Action Plan for Adverse Drug Event Prevention. Washington, DC: US Department of Health and Human Services; 2014. https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf. Accessed June 11, 2018.

2. Centers for Medicare & Medicaid Services. Medicare Part D Drugs. Baltimore, MD: US Department of Health and Human Services; 2017. https://portal.cms.gov/wps/portal/unauthportal/unauthmicrostrategyreportslink?evt=2048001&src=mstrWeb.2048001&documentID=203D830811E7EBD800000080EF356F31&visMode=0&currentViewMedia=1&Server=E48V126P&Project=OIPDA-BI_Prod&Port=0&connmode=8&ru=1&share=1&hiddensections=header,path,dockTop,dockLeft,footer. Accessed October 16, 2018.

3. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older Americans. N Engl J Med. 2011;365(21):2002-2012. doi: 10.1056/NEJMsa1103053.

4. Carrier M, Le Gal G, Wells PS, Rodger MA. Systematic review: case-fatality rates of recurrent venous thromboembolism and major bleeding events among patients treated for venous thromboembolism. Annals of internal medicine. 2010;152(9):578-589. doi: 10.7326/0003-4819-152-9-201005040-00008.

5. Douketis JD, Arneklev K, Goldhaber SZ, Spandorfer J, Halperin F, Horrow J. Comparison of bleeding in patients with nonvalvular atrial fibrillation treated with Ximelagatran or Warfarin: Assessment of incidence, case-fatality rate, time course and sites of bleeding, and risk factors for bleeding. Arch Intern Med. 2006;166(8):853-859. doi: 10.1001/archinte.166.8.853

6. Linkins LA, Choi PT, Douketis JD. Clinical impact of bleeding in patients taking oral anticoagulant therapy for venous thromboembolism: A meta-analysis. Annals of internal medicine. 2003;139(11):893-900.

7. Hylek EM, Held C, Alexander JH, et al. Major bleeding in patients with atrial fibrillation receiving apixaban or warfarin: The ARISTOTLE Trial (Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation): Predictors, Characteristics, and Clinical Outcomes. Journal of the American College of Cardiology. 2014;63(20):2141-2147. doi: 10.1016/j.jacc.2014.02.549.

8. Hylek EM, Singer DE. Risk factors for intracranial hemorrhage in outpatients taking warfarin. Annals of internal medicine. 1994;120(11):897-902.

9. Punthakee X, Doobay J, Anand SS. Oral-anticoagulant-related intracerebral hemorrhage. Thromb Res. 2002;108(1):31-36.

10. Holbrook A, Schulman S, Witt DM, et al. Evidence-based management of anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e152S-184S. doi: 10.1378/chest.11-2295.

11. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology. 2014;64(21):e1-76. doi: 10.1016/j.jacc.2014.03.022.

12. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. Chest. 2013;143(3):751-757. doi: 10.1378/chest.12-1119.

13. Rose AJ, Park A, Gillespie C, et al. Results of a regional effort to improve warfarin management. Annals of Pharmacotherapy. 2017. doi: 10.1177/1060028016681030.

14. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. Thromb Res. 2013;132(2):e124-130. doi: 10.1016/j.thromres.2013.06.006.

15. Nelson WW, Wang L, Baser O, Damaraju CV, Schein JR. Out-of-range international normalized ratio values and healthcare cost among new warfarin patients with non-valvular atrial fibrillation. Journal of medical economics. 2015;18(5):333-340. doi: 10.3111/13696998.2014.1001851.

16. Deitelzweig S, Evans M, Hillson E, et al. Warfarin time in therapeutic range and its impact on healthcare resource utilization and costs among patients with nonvalvular atrial fibrillation. Curr Med Res Opin. 2016;32(1):87-94. doi: 10.1185/03007995.2015.1103217.

17. Liu S, Li X, Shi Q, et al. Outcomes associated with warfarin time in therapeutic range among US veterans with nonvalvular atrial fibrillation. Curr Med Res Opin. 2018;34(3):415-421. doi: 10.1080/03007995.2017.1384370.

18. Labaf A, Sjalander A, Stagmo M, Svensson PJ. INR variability and outcomes in patients with mechanical heart valve prosthesis. Thromb Res. 2015;136(6):1211-1215. doi: 10.1016/j.thromres.2015.10.044.

19. Vanerio G. International Normalized Ratio Variability: A Measure of Anticoagulation Quality or a Powerful Mortality Predictor. J Stroke Cerebrovasc Dis. 2015;24(10):2223-2228. doi: 10.1016/j.jstrokecerebrovasdis.2015.05.017.

20. National Quality Forum. Measure Information: #0555 INR Monitoring for Individuals on Warfarin, Last Updated Jul 02, 2015. 2015.

21. Schein JR, White CM, Nelson WW, Kluger J, Mearns ES, Coleman CI. Vitamin K antagonist use: evidence of the difficulty of achieving and maintaining target INR range and subsequent consequences. Thromb J. 2016;14:14. doi: 10.1186/s12959-016-0088-y.

S.4. Numerator Statement: The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.

S.6. Denominator Statement: Continuously enrolled individuals, at least 18 years of age at of the end of the measurement period, with at least 56 days of warfarin therapy during the measurement period.

S.8. Denominator Exclusions: 1. Individuals who are monitoring INR at home. These individuals are excluded because the claims associated with home INR monitoring are associated with up to four INR tests per claim. Therefore, a single claim for home INR monitoring would not be representative of a single INR test and would prohibit being able to distinguish if the home INR test was within the 56-day timeframe specified by the numerator of this measure.

2. Individuals who have first or last warfarin claims with missing days' supply.

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 05, 2009 **Most Recent Endorsement Date:** Nov 10, 2014

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? Not applicable

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

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

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

NQF_0555_Measure_Evidence_Attachment_-_Final_181029-636764172797235295.docx

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

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

Yes

1a. Evidence (subcriterion 1a)

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

☐ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: INR Monitoring for Individuals on Warfarin

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Prior Submission:

An important consideration for avoiding bleeding and thromboembolic events in patients on warfarin therapy is maintaining the patient's International Normalized Ratio (INR) within the therapeutic range through appropriate and timely INR monitoring and dose adjustment. The recommended range of INR values is 2 to 3 for most conditions treated with warfarin, including deep venous thrombosis, pulmonary embolus, tissue heart valves, atrial fibrillation, and recurrent systemic embolism (Holbrook et al., 2012). The authors of the 2012 American College of Chest Physicians guidelines for antithrombotic therapy and prevention of thrombosis recommend INR monitoring frequency of up to 12 weeks for patients with stable INRs (Holbrook et al., 2012). The latest American College of Cardiology/American Heart Association guidelines continue to recommend INR monitoring on a monthly basis for patients with atrial fibrillation when anticoagulation is stable (Anderson et al., 2013). This measure adopts a conservative approach to INR monitoring of individuals on warfarin by using a

56-day interval, chosen "because a gap of 56 days is traditionally understood to indicate a lack of monitoring, and a period across which TTR is not interpolated" (Rose et al., 2013).

The measure focus is on establishing a minimal INR monitoring interval for the majority of patients on warfarin in the measure denominator. Warfarin has a narrow therapeutic range and therefore, requires regular monitoring with the International Normalized Ratio (INR) test and dose adjustment for the patient to stay within the therapeutic range and avoid thromboembolism or bleeding complications.

Links of Process → Health Outcome

Regular monitoring of patients on warfarin with the International Normalized Ratio (INR) test → More time within the therapeutic range of warfarin → Fewer bleeding and thromboembolic events → Lower hospitalization rates and lower mortality rates

Summary

The desired outcome for this measure is fewer bleeding and thromboembolic events in individuals on warfarin. More regular INR monitoring of patients on warfarin should result in more time in the therapeutic range, resulting in fewer bleeding and thromboembolic events and thus, fewer hospitalizations and deaths.

Citations for 1a.3

Anderson, J. L., Halperin, J. L., Albert, N. M., Bozkurt, B., Brindis, R. G., Curtis, L. H., . . . Shen, W. K. (2013). Management of patients with atrial fibrillation (Compilation of 2006 ACCF/AHA/ESC and 2011 ACCF/AHA/HRS recommendations): A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation, 127*, 1916-1926.

Holbrook, A., Schulman, S., Witt, D. M., Vandvik, P. O., Fish, J., Kovacs, M. J., . . . Guyatt, G. H. (2012). Evidence-based management of anticoagulant therapy: Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest, 141*(2), e152S-e184S.

Rose, A. J., Miller, D. R., Ozonoff, A., Berlowitz, D. R., Ash, A. S., Zhao, S., . . . Hylek, E. M. (2013). Gaps in monitoring during oral anticoagulation: Insights into care transitions, monitoring barriers, and medication nonadherence. *Chest, 143*(3), 751-757.

Updated Evidence:

The primary indications for warfarin are prophylaxis and treatment of venous thromboembolism (VTE) and thromboembolic complications associated with atrial fibrillation.[1] Clinical practice guidelines recommend regular INR monitoring for patients taking warfarin. The 2012 clinical practice guideline from the American College of Chest Physicians (CHEST) suggests an INR testing interval of up to 12 weeks for patients with consistently stable INRs.[2] For patients with previously stable therapeutic INRs who present with a single out-of-range INR of less than or equal to 0.5 below or above therapeutic, the CHEST guideline suggests continuing the current dose and testing the INR within 1 to 2 weeks. In the 2014 guideline from the American College of Cardiology/American Heart Association Task Force for the management of patients with atrial fibrillation, patients are recommended to have INR testing at least monthly when anticoagulation is stable.[3]

Although the guidelines all note that regular INR monitoring is required, the recommended interval for monitoring varies. The CHEST guideline specifically states that "the appropriate length of the recall interval depends on the duration of prior stability and foreseeable future changes in medications or disorders that affect the INR."[2] Studies have shown that maintaining INR stability is challenging. One study found that among patients who reached time in therapeutic range (TTR) ≥ 80% during the first six months, only 42% maintained TTR ≥ 80% during the subsequent 12 months.[4] Another study based on a Canadian cohort noted similar results in examining patients who achieved a TTR > 65% in the first six months of warfarin therapy; only about half remained on warfarin and continued to have good control (TTR > 65%) for months 7 to 12.[5] Given the difficulty in maintaining optimal TTR, the majority of patients taking warfarin are not likely suitable candidates for extended 12-week INR monitoring. Therefore, this measure continues to adopt a conservative approach to INR monitoring of individuals on warfarin by using a 56-day interval (i.e., one INR testing at least every 8 weeks).

The 56-day interval is chosen based on evidence linking to INR control without the burden of excessive testing placed on providers and patients. A large study conducted with 56,490 patients in the Veterans Health Administration (VA) demonstrated a link between gaps in the INR monitoring interval of greater than 56 days and a decrease in TTR. At the patient level, TTR for patients with ≥2 gaps per year was 10 percentage points lower than patients without gaps. At the facility-level, for each gap per patient-year, there was an associated 9.2 percentage point decrease in the facility-level TTR (p<0.001).[6] The monitoring interval of 6 to 8 weeks has also been demonstrated to provide similar INR control as the 4-week interval. A study of 890 patients from six anticoagulant clinics found that the proportion of out-of-range INR results is comparable between patients with and without extended interval monitoring (27.3% vs. 28.4%, p=0.46). The same observation was noted for the extreme out-of-range INR (≤ 1.5 or ≥ 4.0), which was 6.4% vs. 7.7% (p=0.11).[7] .

The linkage between the 56-day monitoring interval and INR control is important because INR variability and TTR are associated with clinical outcomes and healthcare resource utilization. A recent study of 127,385 US veterans provides support for this process-outcome linkage: "Patients with TTR <65% had a higher risk for any stroke/SE [systemic embolism] (HR: 1.57; 95% CI: 1.41–1.75), major bleeding (HR: 2.78; 95% CI: 2.55–3.03) and all-cause mortality (HR: 1.73; 95% CI: 1.67–1.79)."[8] These findings are similar to another study that found that INR variability was shown to be a predictor of mortality and TTR was correlated with patient survival time.[9,10] Lastly, significantly higher stroke-related healthcare costs[11] and total healthcare costs were associated with patients with low TTR (<60%) than those with high TTR.[11,12]

Given the variation in INR monitoring interval recommendations and the evidence suggesting room for improvement in anticoagulation management, this measure supports anticoagulation management through INR monitoring by specifying an evidenced-based interval of 56 days (8 weeks).

Links of Process → Health Outcome

Regular monitoring of patients on warfarin with the International Normalized Ratio (INR) test → More time within the therapeutic range of warfarin → Fewer bleeding and thromboembolic events → Lower hospitalization rates and lower mortality rates

Citations:

1. Coumadin US Full Prescribing Information. Bristol-Myers Squibb. https://packageinserts.bms.com/pi/pi_coumadin.pdf Accessed August 22, 2018.
2. Holbrook, A., Schulman, S., Witt, D. M., Vandvik, P. O., Fish, J., Kovacs, M. J., Guyatt, G. H. (2012). Evidence-based management of anticoagulant therapy: Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, *141*(2), e152S-e184S.
3. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *Journal of the American College of Cardiology*. 2014;64(21):e1-76. doi: 10.1016/j.jacc.2014.03.022.
4. Pokorney SD, Simon DN, Thomas L, et al. Stability of International Normalized Ratios in Patients Taking Long-term Warfarin Therapy. *JAMA*. 2016;316(6):661-663. doi: 10.1001/jama.2016.9356..
5. McAlister FA, Wiebe N, Hemmelgarn BR. Time in therapeutic range and stability over time for warfarin users in clinical practice: a retrospective cohort study using linked routinely collected health data in Alberta, Canada. *BMJ Open*. 2018;8(1):e016980. doi: 10.1136/bmjopen-2017-016980.
6. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. *Chest*. 2013;143(3):751-757. doi: 10.1378/chest.12-1119.
7. Barnes GD, Kong X, Cole D, et al. Extended International Normalized Ratio testing intervals for warfarin-treated patients. *J Thromb Haemost*. 2018;16(7):1307-1312. doi: 10.1111/jth.14150.

8. Liu S, Li X, Shi Q, et al. Outcomes associated with warfarin time in therapeutic range among US veterans with nonvalvular atrial fibrillation. *Curr Med Res Opin.* 2018;34(3):415-421. doi: 10.1080/03007995.2017.1384370.
9. Vanerio G. International Normalized Ratio Variability: A Measure of Anticoagulation Quality or a Powerful Mortality Predictor. *J Stroke Cerebrovasc Dis.* 2015;24(10):2223-2228. doi: 10.1016/j.jstrokecerebrovasdis.2015.05.017.
10. Labaf A, Sjalander A, Stagmo M, Svensson PJ. INR variability and outcomes in patients with mechanical heart valve prosthesis. *Thromb Res.* 2015;136(6):1211-1215. doi: 10.1016/j.thromres.2015.10.044.
11. Deitelzweig S, Evans M, Hillson E, et al. Warfarin time in therapeutic range and its impact on healthcare resource utilization and costs among patients with nonvalvular atrial fibrillation. *Curr Med Res Opin.* 2016;32(1):87-94. doi: 10.1185/03007995.2015.1103217.
12. Nelson WW, Wang L, Baser O, Damaraju CV, Schein JR. Out-of-range international normalized ratio values and healthcare cost among new warfarin patients with non-valvular atrial fibrillation. *Journal of medical economics.* 2015;18(5):333-340. doi: 10.3111/13696998.2014.1001851.

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

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

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

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for intermediate outcome, PROCESS, or STRUCTURE PERFORMANCE measures, including those that are instrument-based) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

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

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Prior Submission:

When the measure was last submitted, some of the questions in the table below were not included in the evidence form. The current submission includes the addition of this information into the evidence form. All updates are presented below in red text.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: Management of patients with atrial fibrillation (Compilation of 2006 ACCF/AHA/ESC and 2011 ACCF/AHA/HRS recommendations): A report of the American College of Cardiology/American Heart Associations Task Force on Practice Guidelines** **Authors: Jeffrey L. Anderson, Jonathan L. Halperin, Nancy M. Albert, Biykem Bozkurt, Ralph G. Brindis, Lesley H. Curtis, David DeMets, Robert A. Guyton, Judith S. Hochman, Richard J. Kovacs, E. Magnus Ohman, Susan J. Pressler, Frank W. Sellke, Win-Kuang Shen** **Date: May 6, 2013** **Citation: Anderson JL, Halperin JL, Albert NM, et al. Management of patients with atrial fibrillation (compilation of 2006 ACCF/AHA/ESC and 2011 ACCF/AHA/HRS recommendations): a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology.* 2013;61(18):1935-1944. doi: 10.1016/j.jacc.2013.02.001. (page 1918)** **URL: https://www.ncbi.nlm.nih.gov/pubmed/23558044**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"5. INR should be determined at least weekly during initiation of therapy and monthly when anticoagulation is stable."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	(definitions provided in Fuster et al., 2011): Class I = Benefit >>> Risk. Procedure/Treatment SHOULD be performed/administered. Evidence Level: A - Multiple populations evaluated. Data derived from multiple randomized clinical trials or meta-analyses.
Provide all other grades and definitions from the evidence grading system	(definitions provided in Fuster et al., 2011): Level B = Limited populations evaluated. Data derived from a single randomized trial or nonrandomized studies. Level C = Very limited populations evaluated. Only consensus opinion of experts, case studies, or standard of care.
Grade assigned to the **recommendation** with definition of the grade	(definitions provided in Fuster et al., 2011): Recommendation: Class I - Multiple populations evaluated. Data derived from multiple randomized clinical trials or meta-analyses.

Provide all other grades and definitions from the recommendation grading system	(definitions provided in Fuster et al., 2011): Class IIa = Benefit >> Risk. Additional studies with focused objectives needed. IT IS REASONABLE to perform procedure/administer treatment. Class IIb = Benefit ≥ Risk. Additional studies with broad objectives needed; additional registry data would be helpful. Procedure/Treatment MAY BE CONSIDERED. Class III No Benefit = Procedure/Test is not helpful. Treatment has no proven benefit. Class III Harm = Procedure/Test entails excess cost without benefit or is harmful. Treatment is harmful to patients.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies on which the recommendation was made: N/A Quality: Evidence Level: A - Multiple populations evaluated. Data derived from multiple randomized clinical trials or meta-analyses. Methods Notes: "This document is a compilation of the current American College of Cardiology Foundation/American Heart Association (ACCF/AHA) practice guideline recommendations for atrial fibrillation (AF) from the "ACC/AHA/ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation," the "2011 ACCF/AHA/HRS Focused Update on the Management of Patients With Atrial Fibrillation (Updating the 2006 Guideline)", and the "2011 ACCF/AHA/HRS Focused Update on the Management of Patients With Atrial Fibrillation (Update on Dabigatran)." Updated and new recommendations from 2011 are noted and outdated recommendations have been removed. No new evidence was reviewed, and no recommendations included herein are original to this document. The ACCF/AHA Task Force on Practice Guidelines chooses to republish the recommendations in this format to provide the complete set of practice guideline recommendations in a single resource."
Estimates of benefit and consistency across studies	The article did not discuss benefit and consistency across studies related to INR monitoring.
What harms were identified?	The article did not discuss harms related to INR monitoring.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Fuster, V., Ryden, L. E., Cannom, D. S., Crijns, H. J., Curtis, A. B., Ellenbogen, K. A., . . . Wann, L. S. (2011). Management of patients with atrial fibrillation: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines developed in partnership with the European Society of Cardiology and in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society. *Journal of the American College of Cardiology, 57*(11), e101-198. See below (Oake et. al., 2008) for additional citations.

Prior Submission:

When the measure was last submitted, some of the questions within the table below were not included in the evidence form. The current submission includes the addition of this information into the evidence form. All updates are presented below in red text.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: Evidence-based management of anticoagulant therapy: Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines** **Authors: Anne Holbrook, Sam Schulman, Daniel M. Witt, Per Olav Vandvik, Jason Fish, Michael J. Kovacs, Peter J. Svensson, David L. Veenstra, Mark Crowther, and Gordon H. Guyatt** **Date: January 23, 2012** **Citation: Holbrook A, Schulman S, Witt DM, et al. Evidence-based management of anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest.* 2012;141(2 Suppl):e152S-184S. doi: 10.1378/chest.11-2295. (page e153S)** **URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278055/**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"3.1. For patients taking VKA therapy with consistently stable INRs, we suggest an INR testing frequency of up to 12 weeks rather than every 4 weeks"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	(definitions provided in Guyatt et al., 2012): Grade 2B: Weak recommendation, moderate-quality evidence
Provide all other grades and definitions from the evidence grading system	(definitions provided in Guyatt et al., 2012): Grade 1A: Strong recommendation, high-quality evidence Grade 1B: Strong recommendation, moderate-quality evidence Grade 1C: Strong recommendation, low- or very-low-quality evidence Grade 2A: Weak recommendation, high-quality evidence Grade 2C: Weak recommendation, low- or very-low-quality evidence
Grade assigned to the **recommendation** with definition of the grade	(definition provided in Guyatt et al., 2012): Grade 2B: Weak recommendation, moderate-quality evidence

Provide all other grades and definitions from the recommendation grading system	(definitions provided in Guyatt et al., 2012): Grade 1A: Strong recommendation, high-quality evidence Grade 1B: Strong recommendation, moderate-quality evidence Grade 1C: Strong recommendation, low- or very-low-quality evidence Grade 2A: Weak recommendation, high-quality evidence Grade 2C: Weak recommendation, low- or very-low-quality evidence
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies on which the recommendation was made: n=3 Quality: Grade 2B: Weak recommendation, moderate-quality evidence Methods notes: "The methods for the development of this article's recommendations follow those developed for the Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Although we aimed to summarize and use randomized controlled trial (RCT) evidence to inform recommendations for clinicians, we found only lower-quality evidence to address most of our questions. At the onset of our review process, our panel decided to limit the recommendations to questions in which evidence met a minimum threshold for quality: at least one comparative study with ≥ 50 patients per group with contemporaneous or historical controls reporting on patient-important outcomes or closely related surrogates. Despite this low threshold, evidence was unavailable for several important clinical management questions. When randomized trials were available, confidence in estimates often decreased because of indirectness (surrogate outcomes) and imprecision (wide CIs)."

Estimates of benefit and consistency across studies	"For patients receiving traditional laboratory-based INR monitoring, retrospective studies have found increasing INR recall intervals associated with both increased and decreased time in therapeutic range (TTR). Other observational studies have suggested that for patients who demonstrate a consistent pattern of stable therapeutic INRs, allowing INR recall intervals of up to 8 weeks would not result in increased risk for bleeding or thromboembolism. Three RCTs have evaluated the effectiveness of INR recall intervals exceeding the traditional North American standard of 4 weeks. One study compared 6- to 4-week recall intervals, whereas another evaluated a flexible approach that allowed recall intervals of up to 12 weeks based on several factors, including the number of prior INRs, longitudinal INR variability, and the risk of adverse events expressed as a function of the INR. The third study compared 4- to 12-week recall intervals using a blinded design. None of the studies found a difference in rates of thromboembolism, bleeding, or INR control. The appropriate length of the recall interval depends on the duration of prior stability and foreseeable future changes in medications or disorders that affect the INR."
What harms were identified?	"None of the studies found a difference in rates of thromboembolism, bleeding, or INR control. The appropriate length of the recall interval depends on the duration of prior stability and foreseeable future changes in medications or disorders that affect the INR."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Guyatt, G. H., Norris, S. L., Schulman, S., Hirsh, J., Eckman, M. H., Akl, E. A., . . . Schünemann, H. J. (2012). Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed.: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest,141*(2_suppl), 53S-70S. See below (Oake et. al., 2008) for additional citations.

Prior Submission:

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: Anticoagulation intensity and outcomes among patients prescribed oral anticoagulant therapy: A systematic review and meta-analysis.** **Authors: Natalie Oake, Alison Jennings, Alan J. Forster, Dean Fergusson, Steve Doucette, & Carl van Walraven** **Date: July 29, 2008** **Citation: Oake N, Jennings A, Forster AJ, Fergusson D, Doucette S, van Walraven C. Anticoagulation intensity and outcomes among patients prescribed oral anticoagulant therapy: a systematic review and meta-analysis. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne.* 2008;179(3):235-244. doi: 10.1503/cmaj.080171.** **URL: http://www.cmaj.ca/content/179/3/235.long**

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	The authors examined evidence as to the risk of hemorrhagic (bleeding) and thromboembolic outcomes associated with levels of the international normalized ratio (INR) that are above and below the recommended range of 2-3. The meta-analysis used person-year data to calculate the relative risk (RR) of hemorrhage and/or thromboembolic event with an INR between 3 and 5, above 5, and below 2, compared to a reference group in the range of 2-3. Effect on Hemorrhagic Events The meta-analysis found that the relative risk of a hemorrhagic event was 2.7 (95% CI: 1.8-3.9) for patients with an INR between 3 and 5 and 21.8 (95% CI: 12.1-39.4) for patients with an INR above 5, compared to those with an INR of 2-3 (the reference group). The relative risks for patients with an INR between 3 and 5 and above 5 were statistically significantly different from those with an INR of 2-3. The above relative risks "translated to absolute risks (and 95% CIs) of 3.7% [per year] (2.2% - 6.3%) for INRs between 3 and 5 and 30.1% [per year] (14.9% - 60.9%) for INRs above 5." These absolute risks can be compared to an absolute risk of 1.4% per year (0.9% - 2.3%) for INRs between 2 and 3. For hemorrhagic events, the relative risks for patients with an INR between 3 and 5, ranged from 0.5 to 11.1 across the 17 individual studies, and for patients with an INR above 5, the relative risks ranged from 4.0 to 161.3. Again, all relative risks are in relation to an INR between 2 and 3. Three studies did not report relative risks for patients with an INR above 5. Effect on Thromboembolic Events The meta-analysis found that the relative risk of a thromboembolic event was 3.5 (95% CI: 2.8–4.4; p<0.01) for patients with an INR less than 2, and 2.6 (95% CI: 1.3–5.1; p<0.01) for patients with an INR above 5, compared to those with an INR of 2-3 (the reference group). The relative risks for patients with an INR less than 2 and above 5 were statistically significantly different from those with an INR of 2-3. These relative risks represent absolute risks (and 95% CIs) of 9.0% per year (6.1% - 13.4%) for INRs less than 2 and 6.6% per year (3.2% -13.9%) for INRs above 5. These absolute risks can be compared to an absolute risk of 2.6% per year (1.8% - 3.6%) for INRs between 2 and 3. For thromboembolic events, the relative risks for patients with an INR less than 2 ranged from 0.0 to 10.9 across the 17 individual studies, and for patients with an INR above 5, the relative risks ranged from 0.0 to 9.0. Again, all relative risks are in relation to an INR between 2 and 3. Six studies did not report relative risks for patients with an INR above 5.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	There was no grade assigned for the quality of quoted evidence.

Provide all other grades and definitions from the evidence grading system	Because there was no grade assigned for the quality of quoted evidence, this information is not available.
Grade assigned to the **recommendation** with definition of the grade	There was no grade assigned for the quality of quoted evidence.
Provide all other grades and definitions from the recommendation grading system	Because there was no grade assigned for the quality of quoted evidence, this information is not available.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies on which the recommendation was made: n=19 Quality: not described. Methods notes: Of the 19 studies included in the systematic review, 10 were retrospective cohort studies, six were randomized controlled trials, and three were prospective cohort studies.

Estimates of benefit and consistency across studies	Effect on Hemorrhagic Events The meta-analysis found that the relative risk of a hemorrhagic event was 2.7 (95% CI: 1.8-3.9) for patients with an INR between 3 and 5 and 21.8 (95% CI: 12.1-39.4) for patients with an INR above 5, compared to those with an INR of 2-3 (the reference group). The relative risks for patients with an INR between 3 and 5 and above 5 were statistically significantly different from those with an INR of 2-3. The above relative risks "translated to absolute risks (and 95% CIs) of 3.7% [per year] (2.2% - 6.3%) for INRs between 3 and 5 and 30.1% [per year] (14.9% - 60.9%) for INRs above 5." These absolute risks can be compared to an absolute risk of 1.4% per year (0.9% - 2.3%) for INRs between 2 and 3. For hemorrhagic events, the relative risks for patients with an INR between 3 and 5, ranged from 0.5 to 11.1 across the 17 individual studies, and for patients with an INR above 5, the relative risks ranged from 4.0 to 161.3. Again, all relative risks are in relation to an INR between 2 and 3. Three studies did not report relative risks for patients with an INR above 5. Effect on Thromboembolic Events The meta-analysis found that the relative risk of a thromboembolic event was 3.5 (95% CI: 2.8–4.4; p<0.01) for patients with an INR less than 2, and 2.6 (95% CI: 1.3–5.1; p<0.01) for patients with an INR above 5, compared to those with an INR of 2-3 (the reference group). The relative risks for patients with an INR less than 2 and above 5 were statistically significantly different from those with an INR of 2-3. These relative risks represent absolute risks (and 95% CIs) of 9.0% per year (6.1% - 13.4%) for INRs less than 2 and 6.6% per year (3.2% -13.9%) for INRs above 5. These absolute risks can be compared to an absolute risk of 2.6% per year (1.8% - 3.6%) for INRs between 2 and 3. For thromboembolic events, the relative risks for patients with an INR less than 2 ranged from 0.0 to 10.9 across the 17 individual studies, and for patients with an INR above 5, the relative risks ranged from 0.0 to 9.0. Again, all relative risks are in relation to an INR between 2 and 3. Six studies did not report relative risks for patients with an INR above 5.
What harms were identified?	Monitoring INR values and titrating warfarin therapy only requires drawing blood and patient counseling and therefore is not generally associated with harms.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The systematic review concluded, "The risks of hemorrhage and thromboemboli are minimized at international normalized ratios of 2–3. Ratios that are moderately higher than this therapeutic range appear safe and more effective than subtherapeutic ratios." Since the publication of the systematic review, we identified six additional studies that support the conclusions of the systematic review and additionally provide evidence that the frequency of INR monitoring is associated with both improved intermediate outcomes (i.e., time in the therapeutic range) and increased risk of thromboembolic events. Citations: Gomes T, Mamdani MM, Holbrook AM, Paterson JM, Hellings C, Juurlink DN. Rates of hemorrhage during warfarin therapy for atrial fibrillation. *CMAJ.* 2013;185(2):E121-127. doi: 10.1503/cmaj.121218. Inoue H, Okumura K, Atarashi H, et al. Target international normalized ratio values for preventing thromboembolic and hemorrhagic events in Japanese patients with non-valvular atrial fibrillation: results of the J-RHYTHM Registry. *Circ J.* 2013;77(9):2264-2270. Rose AJ, Ozonoff A, Henault LE, Hylek EM. Warfarin for atrial fibrillation in community-based practise. *J Thromb Haemost.* 2008;6(10):1647-1654. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. *Chest.* 2013;143(3):751-757. doi: 10.1378/chest.12-1119. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. *Thromb Res.* 2013;132(2):e124-130. doi: 10.1016/j.thromres.2013.06.006. Witt DM, Delate T, Clark NP, et al. Twelve-month outcomes and predictors of very stable INR control in prevalent warfarin users. *J Thromb Haemost.* 2010;8(4):744-749. doi: 10.1111/j.1538-7836.2010.03756.x.

Updated Evidence:

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society.** **Authors: Craig T. January, L. Samuel Wann, Joseph S. Alpert, Hugh Calkins, Joaquin E. Cigarroa, Joseph C. Cleveland Jr., Jamie B. Conti, Patrick T. Ellinor, Michael D. Ezekowitz, Michael E. Field, Katherine T. Murray, Ralph L. Sacco, William G. Stevenson, Patrick J. Tchou, Cynthia M. Tracy and Clyde W. Yancy** **Date: December 2, 2014** **Citation: January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *Journal of the American College of Cardiology.* 2014;64(21):e1-76. doi: 10.1016/j.jacc.2014.03.022. *(page 2251)*** **URL: http://www.onlinejacc.org/content/64/21/e1**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	*"6. Among patients treated with warfarin, the INR should be determined at least weekly during initiation of antithrombotic therapy and at least monthly when anticoagulation (INR in range) is stable."*
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Level of Evidence A: Data derived from multiple randomized clinical trials or meta-analyses.
Provide all other grades and definitions from the evidence grading system	Level of Evidence B: Data derived from a single randomized trial, or nonrandomized studies. Level of Evidence C: Consensus opinion of experts, case studies, or standard of care.
Grade assigned to the **recommendation** with definition of the grade	Class I: Conditions for which there is evidence and/or general agreement that a given procedure or treatment is useful and effective.

Provide all other grades and definitions from the recommendation grading system	Class II: Conditions for which there is conflicting evidence and/or a divergence of opinion about the usefulness/efficacy of a procedure or treatment. IIa: Weight of evidence/opinion is in favor of usefulness/efficacy IIb: Usefulness/efficacy is less well established by evidence/opinion. Class III: Conditions for which there is evidence and/or general agreement that the procedure/treatment is not useful/effective and in some cases may be harmful. No Benefit - Procedure/Test not helpful or Treatment without established proven benefit Harm - Procedure/Test leads to excess cost without benefit or is harmful, and or Treatment is harmful
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies on which the recommendation was made: n=3 Quality: Level of Evidence A: Data derived from multiple randomized clinical trials or meta-analyses. Methods notes: "An extensive evidence review was conducted, focusing on 2006 through October 2012 and selected other references through March 2014." "Searches were extended to studies, reviews, and other evidence conducted in human subjects, published in English, and accessible through PubMed, EMBASE, Cochrane, Agency for Healthcare Research and Quality Reports, and other selected databases relevant to this guideline." "Additionally, the writing committee reviewed documents related to atrial fibrillation (AF) previously published by the ACC and AHA. References selected and published in this document are representative and not all-inclusive."
Estimates of benefit and consistency across studies	The article did not discuss benefits or consistency across studies related to INR monitoring.
What harms were identified?	The article did not discuss harms related to INR monitoring.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Witt et al (2016) published guidance for the management of warfarin therapy. The guidance provided was based on a review of medical literature and consensus opinions of all authors and the endorsement of the Anticoagulation Forum's Board of Directors. The guidance (below) supports an INR monitoring interval of up to 12 weeks. This guidance supports the recommendations and does not change the concussions of the systematic review by Holbrook et al. (2012). Further, Witt et al. (2016) shows the ambiguity in the appropriate length for follow-up. • "During the first 3 months of warfarin therapy for VTE we suggest that INR recall intervals not exceed 6 weeks." • "For patients demonstrating consistently stable INRs after 3 months of warfarin therapy for VTE we suggest that INR recall intervals can be extended up to 12 weeks." The additional studies cited below support the recommendations of the presented evidence that INR should be regularly monitored for patients on warfarin. <u>Citations:</u> Barnes GD, Lucas E, Alexander GC, Goldberger ZD. National trends in ambulatory oral anticoagulant use. *The American journal of medicine.* 2015;128(12):1300-1305 e1302. doi: 10.1016/j.amjmed.2015.05.044. Deitelzweig S, Evans M, Hillson E, et al. Warfarin time in therapeutic range and its impact on healthcare resource utilization and costs among patients with nonvalvular atrial fibrillation. *Current medical research and opinion.* 2016;32(1):87-94. doi: 10.1185/03007995.2015.1103217. Hylek EM, Held C, Alexander JH, et al. Major bleeding in patients with atrial fibrillation receiving apixaban or warfarin: The ARISTOTLE Trial (Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation): Predictors, Characteristics, and Clinical Outcomes. *Journal of the American College of Cardiology.* 2014;63(20):2141-2147. doi: 10.1016/j.jacc.2014.02.549. Iung B, Vahanian A. Epidemiology of acquired valvular heart disease. *The Canadian journal of cardiology.* 2014;30(9):962-970. doi: 10.1016/j.cjca.2014.03.022. Nelson WW, Wang L, Baser O, Damaraju CV, Schein JR. Out-of-range international normalized ratio values and healthcare cost among new warfarin patients with non-valvular atrial fibrillation. *Journal of medical economics.* 2015;18(5):333-340. doi: 10.3111/13696998.2014.1001851. Razouki Z, Ozonoff A, Zhao S, Jasuja GK, Rose AJ. Improving quality measurement for anticoagulation: adding international normalized ratio variability to percent time in therapeutic range. *Circulation Cardiovascular quality and outcomes.* 2014;7(5):664-669. doi: 10.1161/CIRCOUTCOMES.114.000804.

	Rose AJ, Park A, Gillespie C, et al. Results of a regional effort to improve warfarin management. *Annals of Pharmacotherapy.* 2017. doi: 10.1177/1060028016681030. Schein JR, White CM, Nelson WW, Kluger J, Mearns ES, Coleman CI. Vitamin K antagonist use: evidence of the difficulty of achieving and maintaining target INR range and subsequent consequences. *Thromb J.* 2016;14:14. doi: 10.1186/s12959-016-0088-y.US Department of Health and Human Services. National action plan for adverse drug event prevention. Washington, DC: US Department of Health & Human Services Office of Disease Prevention and Health Promotion; 2014. http://health.gov/hcq/ade-action-plan.asp. Accessed November 17, 2015.

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

Warfarin remains the most commonly prescribed anticoagulant in the United States overall[1] and among Medicare PDP beneficiaries.[2] Warfarin has a narrow therapeutic range requiring regular monitoring with the INR test and dose adjustment to maintain patient safety by avoiding thromboembolism or bleeding complications. Warfarin has been identified as the leading drug class implicated in emergency hospitalizations for adverse drug events in adults over 65 years of age.[3] Consequences of adverse drug events related to warfarin therapy are serious and can be fatal. One study found a case-fatality rate of 11.3% for venous thromboembolism (VTE).[4] Case fatality rates for patients with major bleeding can range from 8 percent to 11 percent[4-7] and can reach 45 percent to 50 percent for those with intracranial bleeding.[8,9] For patients with stable INRs, clinical practice guidelines recommend frequent and continuous INR monitoring every 4 to 12 weeks.[10,11] This measure aims to promote patient safety through medication management of individuals on warfarin and to encourage providers to conduct regular INR monitoring for these individuals. Regular INR monitoring is associated with increased time in therapeutic range [12-14] and reduced risk of thromboembolism,[14] whereas subtherapeutic INR is correlated with significantly higher total healthcare

costs[15, 16] and greater risks of stroke/SE,[17] major bleeding[17,18], thromboembolism,[18] and mortality.[17-19]

Current health plan-level performance indicates a quality gap remains. Using 2016 QHP claims data, we found there is a 15.2% difference between the 10th and 90th percentiles with a median score of 56.6% indicating that just over half of health plan members receive regular INR monitoring. In 2016 Medicare claims data, there is an 18.2% difference between the 10th and 90th percentiles with a median score of 71.4% among prescription drug plans. This is a decrease in performance over time compared to the measure developer's previous testing information using data from Medicare prescription drug plans from 2012, which showed a median score of 75.6% and percentiles (P) of performance as follows: P10=64.8%, P25=68.5%, P50=75.6%, P75=81.0%, P90=83.6% indicating variation in performance and room for improvement.[20]

Studies from the literature also suggest an opportunity for improvement in the management of patients on warfarin. A 2015 retrospective study of 9,433 patients who received warfarin for >6 months found that 39% of INR values were out of range.[15] A 2016 review of 6 meta-analyses evaluating the stability of INR (i.e., greater than or equal to 65% time in therapeutic range [TTR]) for patients on anticoagulation therapy found that there is high variability among patients and when patients achieve the target INR range, they do not remain stable and typically have INR values below the therapeutic range, increasing their risk of adverse drug events.[21] A study published in 2018 provides support for the process-outcome linkage: "Patients with TTR <65% had a higher risk for any stroke/SE (HR: 1.57; 95% CI: 1.41–1.75), major bleeding (HR: 2.78; 95% CI: 2.55–3.03) and all-cause mortality (HR: 1.73; 95% CI: 1.67–1.79)."[17] These findings are similar to another study that found that INR variability was shown to be a predictor of mortality where patients with more TTR had higher survival time.[19] The association between TTR and thromboembolism, major bleeding, and death has also been demonstrated in a sample of patients with mechanical heart valve prosthesis.[18]

The literature combined with our empirical evidence suggests room for improvement in anticoagulation management which this measure supports through INR monitoring by specifying an evidence-based interval of 56 days (8 weeks).[12] Further, NQF 0555 is the only endorsed measure that addresses regular monitoring for individuals on warfarin. While NQF 0555 is related to both NQF 0556 (INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications) and NQF 2732 (INR Monitoring for Individuals on Warfarin after Hospital Discharge), all three measures have different clinical foci and target populations. These measures are discussed further in question 5a.2

Citations

1. US Department of Health and Human Services, Office of Disease Prevention and Health Promotion. National Action Plan for Adverse Drug Event Prevention. Washington, DC: US Department of Health and Human Services; 2014. https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf. Accessed June 11, 2018.

2. Centers for Medicare & Medicaid Services. Medicare Part D Drugs. Baltimore, MD: US Department of Health and Human Services; 2017. https://portal.cms.gov/wps/portal/unauthportal/unauthmicrostrategyreportslink?evt=2048001&src=mstrWeb.2048001&documentID=203D830811E7EBD800000080EF356F31&visMode=0&currentViewMedia=1&Server=E48V126P&Project=OIPDA-BI_Prod&Port=0&connmode=8&ru=1&share=1&hiddensections=header,path,dockTop,dockLeft,footer. Accessed October 16, 2018.

3. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older Americans. N Engl J Med. 2011;365(21):2002-2012. doi: 10.1056/NEJMsa1103053.

4. Carrier M, Le Gal G, Wells PS, Rodger MA. Systematic review: case-fatality rates of recurrent venous thromboembolism and major bleeding events among patients treated for venous thromboembolism. Annals of internal medicine. 2010;152(9):578-589. doi: 10.7326/0003-4819-152-9-201005040-00008.

5. Douketis JD, Arneklev K, Goldhaber SZ, Spandorfer J, Halperin F, Horrow J. Comparison of bleeding in patients with nonvalvular atrial fibrillation treated with Ximelagatran or Warfarin: Assessment of incidence,

case-fatality rate, time course and sites of bleeding, and risk factors for bleeding. Arch Intern Med. 2006;166(8):853-859. doi: 10.1001/archinte.166.8.853

6. Linkins LA, Choi PT, Douketis JD. Clinical impact of bleeding in patients taking oral anticoagulant therapy for venous thromboembolism: A meta-analysis. Annals of internal medicine. 2003;139(11):893-900.

7. Hylek EM, Held C, Alexander JH, et al. Major bleeding in patients with atrial fibrillation receiving apixaban or warfarin: The ARISTOTLE Trial (Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation): Predictors, Characteristics, and Clinical Outcomes. Journal of the American College of Cardiology. 2014;63(20):2141-2147. doi: 10.1016/j.jacc.2014.02.549.

8. Hylek EM, Singer DE. Risk factors for intracranial hemorrhage in outpatients taking warfarin. Annals of internal medicine. 1994;120(11):897-902.

9. Punthakee X, Doobay J, Anand SS. Oral-anticoagulant-related intracerebral hemorrhage. Thromb Res. 2002;108(1):31-36.

10. Holbrook A, Schulman S, Witt DM, et al. Evidence-based management of anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e152S-184S. doi: 10.1378/chest.11-2295.

11. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology. 2014;64(21):e1-76. doi: 10.1016/j.jacc.2014.03.022.

12. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. Chest. 2013;143(3):751-757. doi: 10.1378/chest.12-1119.

13. Rose AJ, Park A, Gillespie C, et al. Results of a regional effort to improve warfarin management. Annals of Pharmacotherapy. 2017. doi: 10.1177/1060028016681030.

14. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. Thromb Res. 2013;132(2):e124-130. doi: 10.1016/j.thromres.2013.06.006.

15. Nelson WW, Wang L, Baser O, Damaraju CV, Schein JR. Out-of-range international normalized ratio values and healthcare cost among new warfarin patients with non-valvular atrial fibrillation. Journal of medical economics. 2015;18(5):333-340. doi: 10.3111/13696998.2014.1001851.

16. Deitelzweig S, Evans M, Hillson E, et al. Warfarin time in therapeutic range and its impact on healthcare resource utilization and costs among patients with nonvalvular atrial fibrillation. Curr Med Res Opin. 2016;32(1):87-94. doi: 10.1185/03007995.2015.1103217.

17. Liu S, Li X, Shi Q, et al. Outcomes associated with warfarin time in therapeutic range among US veterans with nonvalvular atrial fibrillation. Curr Med Res Opin. 2018;34(3):415-421. doi: 10.1080/03007995.2017.1384370.

18. Labaf A, Sjalander A, Stagmo M, Svensson PJ. INR variability and outcomes in patients with mechanical heart valve prosthesis. Thromb Res. 2015;136(6):1211-1215. doi: 10.1016/j.thromres.2015.10.044.

19. Vanerio G. International Normalized Ratio Variability: A Measure of Anticoagulation Quality or a Powerful Mortality Predictor. J Stroke Cerebrovasc Dis. 2015;24(10):2223-2228. doi: 10.1016/j.jstrokecerebrovasdis.2015.05.017.

20. National Quality Forum. Measure Information: #0555 INR Monitoring for Individuals on Warfarin, Last Updated Jul 02, 2015. 2015.

21. Schein JR, White CM, Nelson WW, Kluger J, Mearns ES, Coleman CI. Vitamin K antagonist use: evidence of the difficulty of achieving and maintaining target INR range and subsequent consequences. Thromb J. 2016;14:14. doi: 10.1186/s12959-016-0088-y.

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.*

Data

Several sources of data were used in testing the measure. Data representing the target population—members enrolled in Affordable Care Act (ACA) Health Insurance Exchange QHP products —are from four issuers, representing seven QHP products in 2015 and eight products in 2016. Patient-level data representing the target population—members enrolled in Affordable Care Act (ACA) Health Insurance Exchange QHP products—were provided to the Measure Developer from one issuer, henceforth Issuer 1. These data were used to calculate all analyses. A data analytic firm provided QHP analytic results for three issuers, henceforth Issuer 2, Issuer 3, and Issuer 4, in lieu of patient-level data. Additionally, national claims data from Medicare Part B and stand-alone Part D prescription drug plans (PDPs) were used to supplement the QHP analyses since limited QHP data were available for testing.

Analytic Processes

Performance scores on the measure as specified are below. To align with the 2018 Quality Rating System Measure Technical Specifications, all analyses included the following analytic processes:[1,2]

- QHP products with 500 or fewer total members were excluded from all analyses, and
- Denominators had to have at least 30 members in order to show the results of analyses.

The 501 member and 30 minimum denominator rules are not part of the measure specifications. The analyses followed these rules to reflect steps that would be taken if the measure were implemented in the Quality Rating System (QHP data). The 501 member and 30 minimum denominator rules were not applied to the Medicare data since the rules are specific to the Quality Rating System (QHP data).

Performance Scores

Overall, across 4 QHP products from 3 QHP Issuers with sufficient denominators to report measure rates, the performance scores ranged from 48.9% to 62.1% in 2015, and from 43.9% to 59.1% in 2016 (see below). In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 71.7%) among Medicare PDPs. The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries who take warfarin therapy.

RESULTS:

QHP Issuer 1, 2015-2016

The issuer data used to calculate the measure represents 289,136 members and 3 QHP products in 2015, and 223,427 members and 3 QHP products in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / B / 419 / 205 / 48.9%

2016 / B / 326 / 143 / 43.9%

QHP Issuer 2, 2015-2016

The issuer data used to calculate the measure represents 1 product with 45,537 members in 2015, and 30,128 members in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / A / 306 / 190 / 62.1%

2016 / A / 203 / 120 / 59.1%

QHP Issuer 3, 2015-2016

The issuer data used to calculate the measure represents 2 products in 2015 representing 14,093 members, and 3 products in 2016 representing 75,637 members.

Year / Product / Denominator / Numerator / Rate

2015 / A / 57 / 32 / 56.1%

2016 / A / 185 / 105 / 56.8%

2015 / B / Insufficient denominator size for calculation

2016 / B / 126 / 71 / 56.3%

Medicare PDPs, 2012*, 2015, 2016

The Medicare data used to calculate the measure includes 1,140,068 beneficiaries in 2015 and 1,059,826 beneficiaries in 2016. Performance scores from the 2012 data are included for comparison; the scores from 2012 reflect the previous measure specifications submitted to NQF for re-endorsement in 2013.

Plans with at least 100 eligible individuals (minimum denominator for reliability of at least 0.7):

Year / n / Mean / Min / Max / STD / IQR / P10 / P25 / P50 / P75 / P90

2012 / 39 / 74.5% / 59.7% / 88.3% / 7.2% /12.6% / 64.8% / 68.5% / 75.6% / 81.0% / 83.6%

2015 / 56 / 76.7% / 42.0% / 89.0% / 7.7% / 7.4% / 68.5% / 73.1% / 77.1% / 80.5% / 87.5%

2016 / 51 / 71.7% / 46.4% / 85.1% / 7.5% / 10.1% / 64.0% / 67.3% / 71.4% / 77.4% / 82.2%

*Results from testing using 2012 data were from the prior submission of this measure. Updated testing results are from 2015 and 2016.

Citations:

1. National Committee for Quality Assurance. HEDIS® 2018 Volume 2 Technical Specifications for Health Plans. Washington, DC: National Committee for Quality Assurance; 2018.

2. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

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.

Not applicable

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.*

Background

Disparities within the QHP data were determined using the limited demographic variables included in our testing data. At this time, information required to calculate certain disparities (e.g., race/ethnicity) is not coded in a standard manner within administrative claims.[1] Further, there is a lack of clarity regarding which entity (e.g., physician, group, plan, and/or employer) is responsible for capturing and reporting these data.[1] Other health plan measures (e.g., HEDIS quality measures) do not currently collect or report quality performance data stratified by sociodemographic factors.[2]

Method

In the disparities analyses for the measure, female is the reference group for gender, white is the reference group for race/ethnicity, the age group 45-64 is the reference for QHP data, 65+ is the reference group for age for Medicare, and Medicare only is the reference group for dual-enrolled status. Results may be interpreted as better, worse, or the same as a reference group.

In order to assess whether disparities in measure performance exist between subpopulations of the measure cohort, we used the method employed by the Agency for Healthcare Research and Quality (AHRQ) for the National Healthcare Quality and Disparities Report. Disparities statistics were only calculated when the comparison and reference denominators both had at least 30 members in the denominator.[1] Disparities between pairs of population groups were considered identified if the following criteria were met:

1. a Z-test for the difference between two proportions, using a pooled estimate of the variance, was significant with an alpha level of less than 0.05,

2. the relative difference between proportions was greater than 10%

P-Value = statistically significant at the alpha <0.05 level two-tailed Z-test)

Relative Difference = [(Comparison group measure score – Reference group measure score) / Reference group measure score] * 100.

Performance scores on the measure as specified are below, stratified by subpopulation. Results are only shown for those that produced results that met the criteria above to be considered a disparity. Overall, the small denominator sizes of the QHP data limited the disparities analyses. Results based on Medicare data are aggregated national measure rates, whereas QHP rates are issuer-product specific.

Results

Among three issuers' QHP products, disparities for sex were not found in either 2015 or 2016. In 2015, in one issuer, and in one product, a disparity by age group was evidenced: the 27 to 44 age group had lower performance compared to the reference group of 45 to 64.

Although statistical significance was found in the results from Medicare PDP data, national measure rates suggest there is not disparity in care between sexes due to a less than 10% relative difference in measure rates in both 2015 and 2016. However, national measure rates among Medicare PDPs suggest that beneficiaries who were younger, did not identify as white, and were dually eligible for Medicare and Medicaid services had lower measure rates.

Overall, the results of disparities analyses support the measurement of the targeted process of care given that disparities were suggested in both QHP and Medicare data.

Issuer 1 – 2015 & 2016: Rates by Age

2015 - Age

A significant relative difference was detected in Product B measure rates between the 27 to 44 age group and the reference age group of 45 to 64 with the younger age group having lower performance. The other two age groups did not have sufficient denominator size for calculation and comparison.

Product / Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

B / 18 - 26 / Insufficient denominator size for calculation

B / 27 - 44 / 37 / 11 / 29.7% / 41.5 / .0073

B / 45 - 64 / 350 / 178 / 50.9% / Reference / Reference

B / 65+ / Insufficient denominator size for calculation

Medicare Part D Prescription Drug Plans – 2015 & 2016: Rates by Demographics

The following displays the demographic characteristics of the denominator and numerator from national 2015 and 2016 Medicare claims data. National measure rates tended to be significantly lower for beneficiaries who were younger, did not identify as white, and were dually eligible for Medicare and Medicaid services.

2015 - Age

Using the 65 and older category as a reference group, there were significant differences in measure rates when compared to each of the other age groups ($p < 0.0001$ for all 3 comparisons). In addition, the relative difference in measure rates were at least 10% higher for those 65 and older compared to all younger age groups, indicating a disparity in INR monitoring by age with younger age groups less likely to be tested.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

18 -26 / 661 / 443 / 67.0%/ -13.52 / .0001

27 - 44 / 15,452 / 9,981 / 64.0%/ -16.65 / .0001

45 - 64 / 107,147 / 70,747 / 66.0%/ -14.80 / .0001

65+ / 1,016,805 / 788,021 / 77.5%/ Reference/ Reference

2016 - Age

Using the 65 and older category as a reference group, there were significant differences in measure rates when compared to each of the other age groups ($p < 0.0001$ for all 3 comparisons). In addition, the relative difference in measure rates were at least 10% higher for those 65 and older compared to all younger age groups, indicating a disparity in INR monitoring by age with younger age groups less likely to be tested.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

18 -26 / 542 / 339 / 62.6%/ -13.6/ .0001

27 - 44 / 13,357 / 7,876 / 59.0%/ -18.5 / .0001

45 - 64 / 96,342 / 59,610 / 61.9%/ -14.5 / .0001

65+ / 949,585 / 687,168 / 72.4%/ Reference/ Reference

2015 - Race

Significant differences exist between all racial categories when comparing to those who identified as white ($p < 0.0001$ for all 4 comparisons). With the exception of the unknown racial category, the relative differences in measure rates were at least 10% with whites having significantly more INR tests than other racial groups, indicating a disparity in INR monitoring by race.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

White / 1,008,019 / 780,361 / 77.4% / Reference/ Reference

African American / 87,155 / 58,985 / 67.7%/ -12.58 / .0001

Hispanic / 13,739 / 8,630 / 62.8%?/ -18.86 / .0001

Other / 23,873 / 15,914 / 66.7%/ -13.89 / .0001

Unknown / 7,282 / 5,215 / 71.6% / -7.49 / .0001

2016 - Race

Significant differences exist between all racial categories when comparing to those who identified as white ($p < 0.0001$ for all 4 comparisons). With the exception of the unknown racial category, the relative differences in measure rates were at least 10% with whites having significantly more INR tests than other racial groups, indicating a disparity in INR monitoring by race.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

White / 937,692 / 679,123 / 72.4%/ Reference/ Reference

African American / 78,690 / 49,534 / 63.0%/ -13.1 / .0001

Hispanic / 12.641 / 7,256 / 57.4%/ -20.7 / .0001

Other / 22,384 / 13,471 / 60.2%/ -16.9 / .0001

Unknown / 8,419 / 5,609 / 66.6%/ -8.0 / .0001

2015 - Dual-Eligible Status

Dual-eligible beneficiaries are those who are eligible for both Medicare and Medicaid due to their percentage of federal poverty level.[4] Significant differences in measure rates were detected between non-dual-eligible and dual-eligible beneficiaries ($p < 0.001$) with non-dual-eligible beneficiaries having a relative difference of more than 10% more INR tests than dual-eligible beneficiaries, indicating a disparity in INR monitoring by dual-eligible status.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

Non-dual-eligible / 890,686 / 696,933 / 78.3% / Reference/ Reference

Dual-eligible / 249,382 / 172,172 / 69.0%/ -11.77 /.0001

2016 - Dual-Eligible Status

Significant differences in measure rates were detected between non-dual-eligible and dual-eligible beneficiaries ($p < 0.001$) with non-dual-eligible beneficiaries having a relative difference of more than 10% more INR tests than dual-eligible beneficiaries, indicating a disparity in INR monitoring by dual-eligible status.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

Non-dual-eligible / 839,127 / 611,798 / 72.9% / Reference/ Reference

Dual-eligible / 220,699 / 143,195 / 64.9%/ -11.0 / .0001

Citations:

1. Escarce JJ, Carreon R, Veselovskiy G, Lawson EH. Collection of race and ethnicity data by health plans has grown substantially, but opportunities remain to expand efforts. Health Aff (Millwood). 2011;30(10):1984-1991. doi: 10.1377/hlthaff.2010.1117.

2. National Committee for Quality Assurance. HEDIS® 2017 Volume 2 Technical Specifications for Health Plans. Washington, DC: National Committee for Quality Assurance; 2017

3. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

4. Centers for Medicare & Medicaid Services. Seniors & Medicare and Medicaid Enrollees. Baltimore, MD: US Department of Health and Human Services; nd. https://www.medicaid.gov/medicaid/eligibility/medicaid-enrollees. Accessed July 27, 2018.

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

Rose et al. (2013)[1] found that 45% of the 56,490 Veterans Health Administration patients included in their study, who were aged 65 years and older, had at least one gap >=56 days in INR monitoring, representing 44,430 total gaps and 4,482,100 days without INR monitoring over the two-year study period. Predictors of any gaps in monitoring during warfarin therapy that were identified in the study included: younger age (age of 65-69 years versus >=75 years [OR: 1.07; 95%CI: 1.01-1.13]), non-white race (non-Hispanic black race [OR: 1.26; 95%CI: 1.14-1.50], Hispanic race [OR: 1.31; 95%CI: 1.14-1.50], and Native American race [OR: 1.32; 95%CI: 1.01-1.73]), and residence in a zip code with a poverty level below the federal poverty line (poverty level 17.8%-100.0% [OR: 1.24; 95%CI: 1.06-1.45]). The findings from this study are consistent with our analyses of Medicare PDP data that suggest non-dual-eligibles, whites, and older adults have significantly more INR testing compared to dual-eligibles, other racial groups, and younger age groups.

Witt et al. (2013) compared 2,544 patients nonadherent to INR monitoring (>=2 missed INR tests in a row) and 4,995 patients adherent to INR monitoring (never missed >=2 INR tests in a row) from Kaiser Permanente Colorado and described patient characteristics associated with INR monitoring nonadherence.[2] The study

found that factors associated with nonadherence to INR testing included: younger age (increasing age [per year] OR: 0.96; 95% CI: 0.95-0.97), and male sex (female OR: 0.85; 95%CI: 0.77-0.95). The findings from this study are consistent with our analyses of Medicare PDP data that suggest that older adults have significantly more INR testing compared to younger age groups; however, our analyses did not indicate any disparities by sex based on the two criteria used to define disparities (i.e., significant difference and >10% relative difference).

Citations

1. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. Chest. 2013;143(3):751-757. doi: 10.1378/chest.12-1119.

2. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. Thromb Res. 2013;132(2):e124-130. doi: 10.1016/j.thromres.2013.06.006.

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):*

Cardiovascular

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

Safety, Safety : Medication

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

Elderly, Populations at Risk

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.)*

Not applicable

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** 0555_INR_CompleteCoding-636764172796610581.xlsx

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

No, this is not an instrument-based measure **Attachment:**

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.

Not an instrument-based measure

S.3.1. <u>For maintenance of endorsement:</u> 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. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Age specification was changed from at least 18 years of age at the beginning of the measurement period to at least 18 years of age as of the end of the measurement period for the purpose of alignment and harmonization with Healthcare Effectiveness Data and Information Set (HEDIS) quality measures.

National Drug Codes (NDCs) have been updated to include new drugs on the market that are applicable to the measure. Drugs that have been discontinued for more than three years have been removed.

Enrollment criteria were changed from enrollment for 11 out of 12 months to enrollment in a Qualified Health Plan (QHP) product for at least two months, with no gap in enrollment between the first enrolled month and last enrolled month of a calendar year. This was done for two reasons: 1) at least two consecutive months are necessary to create a 56-day interval and 2) to maximize the number of patients eligible for the measure. The latter rationale adapts the measure for member turnover within QHP products operating in the Health Insurance Exchange. Utilizing the previous specifications of enrollment for 11 out of 12 months resulted in approximately 50% of the members in our QHP sample that would not meet the criteria to be included in the measure.

The following describes the terminology of the units associated with the Health Insurance Exchange used throughout this form: “Issuer” refers to an individual insurance company or insurance organization. The term “product” refers to a package of health coverage benefits that are offered using a particular network type (i.e., health maintenance organization, preferred provider organization, exclusive provider organization, point of service, or indemnity). Unique products for each issuer are referred to using alphabetic labeling (e.g., two unique products from the same issuer are referred to as Product A and Product B).

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.*

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

The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.

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)*

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

Individuals in the denominator who have at least one INR test performed during each 56-day interval with warfarin therapy will be counted in the numerator. All 56-day intervals in which an individual is both prescribed warfarin and continuously enrolled are used to calculate the INR compliance rate for the individual. A 56-day interval with a hospitalization of more than 48 hours is considered an interval with an INR test.

Interval: The first day of the first 56-day interval is the start date of the first warfarin prescription in the measurement period, and the last day of the first 56-day interval is the start date of the first warfarin

prescription + 55 days. The subsequent 56-day interval starts on the day after the first 56-day interval and ends 56 days following the first 56-day interval, as long as this end date occurs within the warfarin therapy time frame. This process continues until a calculated 56-day interval end date does not occur within the warfarin therapy time frame. If there are fewer than 56 days of warfarin therapy within the warfarin therapy time frame, those remaining days are not counted in any interval in determining the numerator. Only full 56-day intervals are used for calculating the numerator. "Warfarin usage" or "warfarin therapy" is determined by the start date of the first prescription for warfarin up through the start date of the last prescription for warfarin plus the days' supply from the last claim.

2015-2017 CODES FOR INR TEST

The specific year of codes used for the measure is dependent upon the measurement year.

CPT code:

85610 – Prothrombin time

LOINC codes:

34714-6 – INR in blood by coagulation assay

5894-1 – Prothrombin time (PT) actual/normal

6301-6 – INR in platelet poor plasma by coagulation assay

38875-1 – INR in platelet poor plasma or blood by coagulation assay

5964-2 – Prothrombin time (PT) in blood by coagulation

5902-2 – Prothrombin time (PT)

6418-0 – INR in capillary blood by coagulation assay [2016 only]

46418-0 – INR in capillary blood by coagulation assay [2017 only]

46417-2 – Prothrombin time (PT) in capillary blood by coagulation assay

52129-4 – INR in platelet poor plasma by coagulation assay—post heparin adsorption

Note: A full list of codes necessary for measure calculation is provided in the attached Excel file.

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

Continuously enrolled individuals, at least 18 years of age at of the end of the measurement period, with at least 56 days of warfarin therapy during the measurement period.

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.)*

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

The time period of the data is defined as any time during the measurement period (12 consecutive months). "Continuously enrolled" for this measure is defined as enrollment in a QHP product for at least two months, with no gap in enrollment between the first enrolled month and last enrolled month of a calendar year. "Warfarin usage" or "warfarin therapy" is determined by the start date of the first prescription for warfarin through the start date of the last prescription for warfarin plus the days' supply from the last claim.

ENROLLMENT CRITERIA

Criteria for QHP products: At least two months enrollment in a QHP product, with no gap in enrollment between the first enrolled month and the last enrolled month of a calendar year.

MEDICATION ACTIVE INGREDIENTS

Active Ingredients by Class: Anticoagulants – Warfarin. Note the active ingredient is limited to oral formulations only. A full list of codes necessary for measure calculation is provided in an attached Excel file.

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

1. Individuals who are monitoring INR at home. These individuals are excluded because the claims associated with home INR monitoring are associated with up to four INR tests per claim. Therefore, a single claim for home INR monitoring would not be representative of a single INR test and would prohibit being able to distinguish if the home INR test was within the 56-day timeframe specified by the numerator of this measure.

2. Individuals who have first or last warfarin claims with missing days' supply.

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.)*

2015-2017 INR MONITORING AT HOME HCPCS CODES:

G0248 – Demonstrate Use Home INR Mon

G0249 – Provide Test Mats & Equip Home INR

G0250 – MD INR Test Review Inter Mgmt

Note: A full list of codes necessary for measure calculation is provided in the attached Excel file.

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

Not applicable

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

No risk adjustment or risk stratification

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.)*

Denominator: Continuously enrolled individuals, at least 18 years of age at of the end of the measurement period, with at least 56 days of warfarin therapy during the measurement period.

Create Denominator:

1. Pull individuals who are at least 18 years of age as of the end of the measurement period.
2. Include individuals who meet continuous enrollment criteria as described above in S.7.
3. Of the individuals identified in Step 2, include those who had warfarin claims during the measurement period.

4. Exclude individuals who have warfarin claims with missing days' supply. Exclude individuals who are monitoring their INR at home.

5. Of the individuals who were not excluded in Step 4, calculate the start date and end date of warfarin therapy for each individual and count the days between the start date and the end date inclusive. If an individual's death date is available, then use the death date as the end date.

6. Keep individuals who had at least 56 days of warfarin therapy during the measurement period and calculate the number of full 56-day intervals for each individual.

Numerator: The number of individuals in the denominator who receive at least one INR monitoring test during each 56-day interval with active warfarin therapy.

Create Numerator:

7. Pull all INR test claims from claims data for the current measurement period.

8. From the claims identified in Step 7, keep only those INR test claims for the individuals who are included in the denominator.

9. From claims data, identify and pull all inpatient stays of more than 48 hours during the measurement period (where hours are not available, calculate and keep stays of at least three days).

10. From the claims identified in Step 9, keep those that are for the individuals who are included in the denominator.

11. Combine the INR test claims dataset from Step 8 and the hospitalizations of more than 48 hours dataset from Step 10.

12. Using the start date of warfarin therapy identified in the denominator, determine the subsequent start dates for each of the calculated 56-day interval(s) of warfarin therapy and determine the number of full 56-day intervals designated in the denominator for each individual.

13. From the dataset created in Step 11, create a dataset containing INR tests performed and inpatient stays by unique individual and date of service.

14. Determine which full 56-day intervals have an INR test completed or have an inpatient stay by comparing each date of service from Step 13 to each full 56-day interval for each individual designated in Step 12.

15. From the dataset created in Step 14, calculate the individual's INR monitoring compliance rate as the sum of the number of full 56-day intervals with an INR test divided by the total number of full 56-day intervals.

16. From the dataset created in Step 15, calculate the measure numerator by counting the number of individuals with a 100% INR monitoring compliance rate.

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.

This measure is not based on a sample.

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.

This measure is not based on survey or patient-reported data.

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

If other, please describe in S.18.

Claims

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.

There is no data collection instrument; individual health plans produce administrative claims in the course of providing care to health plan members.

The following sources of data are needed to calculate NQF 0555:

1. QHP products: Claims data from issuers, consisting of hospital and office visits, pharmacy, and laboratory claims (when available); enrollment data; and members' demographic data OR

2. Medicare: Claims data from Medicare Parts A, B and D consisting of inpatient and outpatient claims and prescription drug events; enrollment data; and beneficiaries' demographic data.

Please note that Medicare data were used for measure testing to enhance the measure testing results. At the time this form was completed, CMS does not yet have any plan to add this measure to any quality reporting or value-based purchasing programs for Medicare beneficiaries but may consider these measures for the future. However, this measure is being considered for use in the Quality Rating System for Qualified Health Plans.

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)*

No data collection instrument provided

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

Health Plan

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

Outpatient Services

If other:

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

Not applicable because this is not a composite performance measure.

2. Validity – See attached Measure Testing Submission Form

NQF_0555_Measure_Testing_Form_-_Final_181029-636764172797860443.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.

No - This measure is not risk-adjusted

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

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

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

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

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☒☒ administrative claims	☒ ☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The following specific datasets were used for testing:

PRIOR SUBMISSION

- 2011–2012 Medicare Parts A, B, and D claims data for 10 states (Arizona, Delaware, Florida, Iowa, Indiana, Mississippi, Missouri, Rhode Island, Texas, and Washington)
- 2011 Medicare Parts A, B, and D claims data for 31 ACOs

UPDATED TESTING

- 2015–2016 administrative claims data from four issuers (referred to as QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4), containing a total of seven Health Insurance Exchange Qualified Health Plan (QHP) products in 2015 and eight in 2016. The following describes the terminology of the units associated with the Health Insurance Exchange: “Issuer” refers to an individual insurance company or insurance organization. The term “product” refers to a package of health coverage benefits that are offered using a particular network type (i.e., health maintenance organization, preferred provider organization, exclusive provider organization, point of service, or indemnity).[1] Unique products for each issuer are referred to

using alphabetic labeling (e.g., two unique products from the same issuer are referred to as Product A and Product B).

- 2015–2016 administrative claims data from Medicare Parts A, B, and D for beneficiaries enrolled in stand-alone Part D Prescription Drug Plans (referred to as Medicare PDPs)

Please note that Medicare data were used for measure testing to enhance the measure testing results. At the time this form was completed, CMS does not yet have any plan to add this measure to any quality reporting or value-based purchasing programs for Medicare enrollees but may consider these measures for the future. However, this measure is being considered for use in the Quality Rating System for Qualified Health Plans.

Citation:

1. Centers for Medicare & Medicaid Services. Federal Definitions for Health Insurance Products and Plans. Baltimore, MD: US Department of Health and Human Services; 2016. https://www.cms.gov/CCIIO/Resources/Training-Resources/Downloads/product-vs-plan-ppt.pdf. Accessed June 12, 2018.

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

PRIOR SUBMISSION

- January 1, 2011 – December 31, 2012

UPDATED TESTING

- January 1, 2015 – December 31, 2016

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ ☒ health plan	☒ ☒ health plan
☒ other: State, Accountable Care Organization	☒other: State, Accountable Care Organization

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*)

PRIOR SUBMISSION

Characteristics of the sample for 2011–2012 are summarized in Table 1. All beneficiaries from 10 states (Arizona, Delaware, Florida, Iowa, Indiana, Mississippi, Missouri, Rhode Island, Texas, and Washington) were included in the testing sample. Measured entities included 10 states, 83 Prescription Drug Plans (PDPs), and 26,182 Physician Groups. Fourteen percent of PDPs had fewer than 30 beneficiaries attributed, accounting for less than 0.01% of total beneficiaries attributed to a PDP. Sixty-five percent of physician groups had fewer than 30 beneficiaries attributed. These groups represent 1.2% of the total number of beneficiaries attributed to a physician group.

Table 1. 2011-2012 Sample Characteristics by States, PDPs, and Physician Groups

Characteristics	States n=10	Prescription Drug Plans n=83	Physician Groups n=26,182
Total Number	14,162,440	14,162,440	14,162,440
Total Attributed (%)	14,162,440 (100%)	4,699,420 (33.18%)	4,241,116 (29.95%)
Mean # of Beneficiaries	1,416,244	56,656	194
Median # of Beneficiaries	1,171,694	1,221	10
Min # of Beneficiaries	183,084	1	1
Max # of Beneficiaries	4,098,325	1,102,813	37,977
STD	1,369,273	167,654	907
P10	200,154	8	1
P25	598,022	113	3
P50	1,171,694	1,221	10
P75	1,213,975	38,693	85
P90	3,896,824	121,506	394

A convenience sample of beneficiaries attributed to 31 Accountable Care Organizations (ACOs) was used for testing the measure at the ACO level. Characteristics of the ACO sample for 2011 are summarized in Table 2.

Table 2. 2011 Sample Characteristics for 31 ACOs

Characteristics	ACOs
Total Number	31
Total Beneficiaries	682,036
Mean # of Beneficiaries	22,001
Median # of Beneficiaries	18,622
Min # of Beneficiaries	7,207
Max # of Beneficiaries	61,957
STD	12,001
P10	10,309
P25	13,249
P50	18,622
P75	24,356
P90	35,853

UPDATED TESTING

Characteristics of the data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are summarized in Tables 3a (2015) and 3b (2016). The data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3,

and QHP Issuer 4 included all members with claims associated with the QHP products. To align with the 2018 Quality Rating System, Measure Technical Specifications: [1]

- QHP products with 500 or fewer total members were excluded from all analyses, and
- Denominators had to have at least 30 members in order to show the results of analyses.

The 501 member and 30 minimum denominator rules are not part of the measure specifications. The analyses followed these rules to reflect steps that would be taken if the measure were implemented into the Quality Rating System (QHP data).

The Medicare sample included all beneficiaries from the national Medicare claims database who had at least one month of Part A and Part B coverage and no HMO coverage during the year and who were in a stand-alone Medicare PDP. The 501 member and 30 minimum denominator rules were not applied to the Medicare data since the rules are specific to the Quality Rating System (QHP data).

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

Table 3a. 2015 Sample Characteristics of the Data

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Number of QHP Products or Medicare PDPs	3	1	2	1	66
Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP	289,136	49,137	15,671	3,354	18,894,628
Mean # of Members/ Beneficiaries per Product/PDP	96,378	49,137	7,836	3,354	286,282

Table 3b. 2016 Sample Characteristics of the Data

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Number of QHP Products or Medicare PDPs	3	1	3	1	62
Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP	223,427	33,205	84,255	2,284	19,607,672
Mean # of Members/ Beneficiaries per Product/PDP	74,476	33,205	28,085	2,284	316,253

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*)

PRIOR SUBMISSION

Demographic characteristics of the beneficiaries in the 2011–2012 datasets are shown in Table 4 below.

Table 4. 2011–2012 Demographic Characteristics by State, PDPs, and Physician Groups

Characteristics	State n=10	Prescription Drug Plans n=83	Physician Groups n=26,182
Total Population	14,162,440	4,699,420	4,241,116
Gender			
Female	6,948,546 (49.06%)	2,697,239 (57.40%)	2,482,734 (58.54%)
Male	5,827,374 (41.15%)	1,782,594 (37.93%)	1,710,539 (40.33%)
Unknown	1,386,520 (9.79%)	219,587 (4.67%)	47,843 (1.13%)
Age			
≥65 years	9,949,181 (70.25%)	3,326,257 (70.78%)	3,334,085 (78.61%)
Race			
White/Caucasian	11,086,802 (78.28%)	3,887,785 (82.73%)	3,693,852 (87.10%)
African-American	1,213,508 (8.57%)	460,400 (9.80%)	335,859 (7.92%)
Hispanic	474,632 (3.35%)	195,928 (4.17%)	109,142 (2.57%)
Other	1,387,498 (9.80%)	155,307 (3.30%)	89,041 (2.10%)
Ethnicity			
Hispanic	474,632 (3.35%)	195,928 (4.17%)	109,142 (2.57%)
Non-Hispanic	13,687,808 (96.65%)	4,503,492 (95.83%)	4,131,974 (97.43%)
Medicare and Medicaid Eligibility			
Dual Eligible	2,029,697 (14.33%)	1,339,687 (28.51%)	785,130 (18.51%)
Non-Dual Eligible	12,132,743 (85.67%)	3,359,733 (71.49%)	3,455,986 (81.49%)

Demographic characteristics of the beneficiaries in the ACO dataset are shown in Table 5.

Table 5. 2011 Demographic Characteristics by ACO

Characteristics	ACO Number (%)
Total Population	682,036
Gender	
Female	398,763 (58.47%)
Male	283,273 (41.53%)
Age	

Characteristics	ACO Number (%)
≥65 years	574,224 (84.34%)
Race	
White/Caucasian	574,672 (84.26%)
African-American	46,211 (6.78%)
Hispanic	21,310 (3.12%)
Other	38,181 (5.60%)
Unknown	1,662 (0.24%)
Ethnicity	
Hispanic	21,310 (3.12%)
Non-Hispanic	660,726 (96.88%)
Medicare and Medicaid Eligibility	
Dual Eligible	152,960 (22.43%)
Non-Dual Eligible	529,076 (77.57%)

UPDATED TESTING

Demographic characteristics of members of QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are shown in Tables 6a (2015) and 6b (2016); however, limited demographic variables were available in our testing data. "N/A" in the tables indicates the data were not available.

Table 6a. 2015 Demographic Characteristics of Members of QHP Issuers and Medicare PDPs

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Sample Size	289,136	49,137	15,671	3,354	18,894,628
Sex n (% of Total Sample)*					
Female	150,116 (51.9)	21,399 (43.5)	7,043 (44.9)	1,538 (45.9)	10,413,926 (55.1)
Male	139,020 (48.1)	27,738 (56.5)	8,628 (55.1)	1,816 (54.1)	8,480,702 (44.9)
Age n (% of Total Sample)*					
<18 years	9,584 (3.3)	3,600 (7.3)	1,578 (10.1)	247 (7.4)	132 (0.0)
18–26 years	38,590 (13.4)	3,633 (7.4)	1,640 (10.5)	333 (9.9)	105,869 (0.6)
27–44 years	81,098 (28.0)	12,486 (25.4)	5,671 (36.2)	1,022 (30.5)	911,610 (4.8)

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
45–64 years	152,252 (52.7)	28,965 (59.0)	6,603 (42.1)	1,711 (51.0)	2,958,692 (15.7)
≥65 years	7,612 (2.6)	453 (0.9)	179 (1.1)	41 (1.2)	14,918,325 (79.0)
Race n (% of Total Sample)*					
White/ Caucasian	N/A	N/A	N/A	N/A	15,782,130 (83.5)
African-American	N/A	N/A	N/A	N/A	1,893,242 (10.0)
Hispanic	N/A	N/A	N/A	N/A	383,461 (2.0)
Other	N/A	N/A	N/A	N/A	633,329 (3.4)
Unknown	N/A	N/A	N/A	N/A	202,466 (1.1)

*Numbers in parentheses represent the column percent by demographic characteristic.

Table 6b. 2016 Demographic Characteristics of Members of QHP Issuers and Medicare PDPs

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Sample Size	223,427	33,205	84,255	2,284	19,607,672
Sex n (% of Total Sample)*					
Female	116,111 (52.0)	14,546 (43.8)	38,433 (45.6)	1,027 (45.0)	10,787,561 (55.0)
Male	107,316 (48.0)	18,659 (56.2)	45,822 (54.4)	1,257 (55.0)	8,820,111 (45.0)
Age n (% of Total Sample)*					
<18 years	8,536 (3.8)	3,077 (9.3)	8,618 (10.2)	207 (9.1)	121 (0.0)
18–26 years	27,732 (12.4)	2,445 (7.4)	8,268 (9.8)	236 (10.3)	101,020 (0.5)
27–44 years	58,419 (26.2)	8,584 (25.8)	27,730 (32.9)	724 (31.7)	888,545 (4.5)
45–64 years	121,304 (54.3)	18,756 (56.5)	38,748 (46.0)	1,089 (47.7)	2,942,822 (15.0)

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
≥65 years	7,436 (3.3)	343 (1.0)	891 (1.1)	28 (1.2)	15,675,164 (79.9)
Race n (% of Total Sample)*					
White/ Caucasian	N/A	N/A	N/A	N/A	16,355,081 (83.4)
African-American	N/A	N/A	N/A	N/A	1,920,626 (9.8)
Hispanic	N/A	N/A	N/A	N/A	402,787 (2.1)
Other	N/A	N/A	N/A	N/A	673,534 (3.4)
Unknown	N/A	N/A	N/A	N/A	255,644 (1.3)

*Numbers in parentheses represent the column percent by demographic characteristic.

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.

UPDATED TESTING

The following sources of data were used in testing NQF 0555 at the health plan level:

1. QHP products: claims data from issuers, consisting of hospital and office visit, pharmacy, and laboratory claims (when available); enrollment data; members' demographic data; and provider information.
2. Medicare: claims data from Medicare Parts A and B and stand-alone Part D PDPs, consisting of inpatient and outpatient claims and prescription drug events; enrollment data; members' demographic data; and provider information.

The difference in the data used for the various aspects of testing is shown in Table 7. "X" indicates no data were available.

Table 7. Data Used to Test the Measure

Testing of the Measure	QHP Data	Medicare Data
Development of the Denominator	✓	✓
Development of the Numerator	✓	✓
Data Element Feasibility	✓	✓
Measure Performance Reliability (Signal to Noise)	✓	✓
Calculating Measure Performance	✓	✓
Convergent Validity	X	✓
Exclusion Analyses	✓	✓
Disparities Analyses	✓	✓

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.

UPDATED TESTING

This process measure, NQF 0555, is not risk adjusted and therefore an analysis of social risk factors was not conducted.

2a2. RELIABILITY TESTING

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

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

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

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

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

PRIOR SUBMISSION

The method of reliability testing used and the rationale are described below.

Method of Reliability Testing and Rationale

In order to assess measure precision in the context of the observed variability across measurement units (states, prescription drug plans [serving as a proxy for health plans], Accountable Care Organizations [ACOs]), we utilized the approach proposed by Adams (2009) and Scholle et al. (2008). The rationale for this choice of testing was based on the work on the reliability of provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial published by Adams: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient."

The signal-to-noise ratio was calculated as a function of the variance between measured entities (signal) and the variance within a measured entity (noise). Reliability was estimated using a beta-binomial model. This approach has two basic assumptions:

1. Each measured entity has a true pass rate, p, which varies from group to group; and,
2. The measured entity's score is a binomial random variable conditional on the entities true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual physician group variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across physician groups). In a simulation, Adams showed that differences between physicians started to be seen at reliability of 0.7 and significant differences could be seen at reliability of 0.9. Our rationale was based on Adams' work, and thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between physicians.

Using methodology described by Scholle et al. (2008), reliability estimates were computed separately based on the mean denominator size for physicians within each denominator category. As Scholle described in the

article, the reliability estimate at the mean denominator for each category should reflect "the typical experience of physicians in this population."

Reliability scores were also calculated for state, prescription drug plan (which served as a proxy for health plans), and ACO levels of measurement using the same approach.

Adams, J. L. The reliability of provider profiling: A tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009.

Scholle, S. H., Roski, J., Adams, J. L., Dunn, D. L., Kerr, E. A., Dugan, D. P., et al. (2008). Benchmarking physician performance: Reliability of individual and composite measures. *American Journal of Managed Care*, *14*(12), 833-838.

UPDATED TESTING

Measure score reliability was estimated using a beta-binomial model. For the QHP data, the mean reliability was calculated across QHP products. Reliability estimates for Medicare PDPs were computed by using the methods of minimum denominator and volume categories, described by Scholle et al. (2008).[1] This difference in approach to the data is due to the limited number of available QHP products.

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

We calculated reliability for each QHP product and the mean reliability across QHP products in 2016. Note that QHP Issuer 4 did not have sufficient denominator sizes for analyses and is thus not presented in the results section for reliability, below. Sufficient denominator size for display was defined as 30 members or more in the denominator to align with the 2018 Quality Rating System Measure Technical Specifications.[2]

Minimum Denominator for Reliability, Medicare PDPs

The testing conducted for this comprehensive re-evaluation used the same methods for the 2016 Medicare PDP sample as described above with the exception that we used the method of minimum denominator and volume categories from Scholle et al. instead of the mean denominator.[2] This method assumes that the denominator size in each volume category is equal to the minimum for that category. As such, it provides a more conservative estimate of reliability for each volume category.

Citations:

1. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care.* 2008;14(12):833-838.

2. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

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

PRIOR SUBMISSION

We conducted reliability tests across measurement units, and the results from the state level, including reliability statistics and assessments of adequacy, are provided below.

We concluded that the reliability test was adequate, since all state-level reliability scores were greater than 0.7, indicating that the measure would produce reliable scores at the state level (Table 8).

Table 8. 2011-2012 State Reliability and Assessment of Adequacy for Tests Conducted

State	Measure Rate (Reliability)
AZ	74.62% (0.99)

State	Measure Rate (Reliability)
DE	75.45% (0.99)
FL	74.28% (0.99)
IA	83.19% (0.99)
IN	77.81% (0.99)
MO	76.30% (0.99)
MS	65.53% (0.99)
RI	88.61% (0.99)
TX	64.28% (0.99)
WA	78.39% (0.99)

Using the method of mean denominator and volume categories, a minimum denominator of 100 individuals resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between plans. Furthermore, more than half (52.0%) of the plans with at least one patient attributed (n=75) had at least 100 individuals in the measure denominator and a reliable score (Table 9).

Table 9. 2012 Prescription Drug Plan Reliability and Assessment of Adequacy for Tests Conducted

Year	Min Denominator	# of Plans (% of PDPs with at least 1 individual attributed)	Mean Rate of Plans	Reliability Score
2012	100	39 (52.0%)	74.52%	0.71

UPDATED TESTING

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

Among the QHP products tested, reliability ranged from 0.60 to 0.79 with a mean reliability of 0.70 (Table 10a), which suggests sufficient signal relative to noise to discriminate performance between plans.

Table 10a. 2016 Reliability Among QHP Products with At Least 30 Members in the Denominator

QHP Issuer	Product	Denominator	Numerator	Measure Rate	Variance Within	Variance Between	Reliability Score
Issuer 1	B	326	143	43.9%	7.55	29.19	0.79
Issuer 2	A	203	120	59.1%	11.91	29.19	0.71
Issuer 3	A	185	105	56.8%	13.27	29.19	0.69
Issuer 3	B	126	71	56.4%	19.52	29.19	0.60
Mean							**0.70**

Minimum Denominator for Reliability, Medicare PDPs

Using the method of minimum denominator and volume categories, a minimum of 100 members in the denominator results in an overall reliability score of 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between units of analysis.[1] Of the 61 PDPs in 2016, the

majority (83.6%) of PDPs had at least 100 individuals in the measure denominator, representing a mean performance rate of 71.74% (reliability = 0.70) (Table 10b).

Table 10b. 2016 Medicare PDP Reliability and Assessment of Adequacy for Tests Conducted

Min Denominator	Total # of PDPs	# of PDPs with at Least 100 Individuals	Mean Rate of Plans with at Least 100 Individuals	Reliability Score
100	61	51	71.74%	0.70

Citation:

1. Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling—reliability and risk of misclassification. *N Engl J Med.* 2010;362(11):1014-1021. doi: 10.1056/NEJMsa0906323.

PRIOR SUBMISSION

Using the method of mean denominator and volume categories, a minimum denominator of 50 individuals measured resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between physician groups. Approximately 11% of physician groups with at least one patient attributed (n=6,594) had at least 50 individuals in the measure denominator and a reliable score (Table 11).

Table 11. 2012 Physician Group Reliability and Assessment of Adequacy for Tests Conducted

Year	Min Denominator	# of Physician Groups (% of physician groups with at least 1 individual attributed)	Mean Rate of Physician Groups	Reliability Score
2012	50	739 (11.21%)	75.66%	0.73

Using the method of mean denominator and volume categories, a minimum denominator of 50 individuals resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between ACOs (Table 12). The aforementioned criteria resulted in 100.0% of all ACOs (31 of 31 ACOs) with reliable scores (Table 13).

Table 12. 2011 ACO Reliability and Assessment of Adequacy for Tests Conducted

Year	**Min Denominator**	**# of ACOs (% of ACOs with at least 1 individual attributed)**	**Mean Rate of ACOs**	**Reliability Score**
2011	50	31 (100.0%)	75.34%	0.71

Table 13. 2011 Individual ACO Reliability and Assessment of Adequacy for Tests Conducted

ACO #	Denominator	Measure Rate (Reliability)
1	1,124	85.14% (0.99)
2	650	87.08% (0.98)
3	1,034	80.37% (0.98)
4	443	70.65% (0.95)
5	466	64.38% (0.95)

ACO #	Denominator	Measure Rate (Reliability)
6	657	81.13% (0.97)
7	502	65.54% (0.95)
8	947	63.99% (0.97)
9	848	58.49% (0.97)
10	394	69.80% (0.94)
11	1,451	84.01% (0.99)
12	402	78.61% (0.96)
13	1,022	85.62% (0.97)
14	902	81.71% (0.98)
15	943	83.67% (0.98)
16	1,609	88.56% (0.98)
17	1,175	81.28% (0.99)
18	473	79.28% (0.99)
19	450	78.00% (0.96)
20	815	90.67% (0.96)
21	697	89.53% (0.99)
22	137	69.34% (0.98)
23	1,577	74.51% (0.97)
24	825	69.33% (0.95)
25	440	62.95% (0.93)
26	350	65.71% (0.96)
27	361	62.60% (0.98)
28	448	76.34% (0.95)
29	481	61.12% (0.95)
30	772	84.59% (0.98)
31	2,003	61.66% (0.99)

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?*)

PRIOR SUBMISSION

The results indicated that the measure, as currently specified, was reliable at the state, prescription drug plan, and ACO levels. However, due to sample size issues only a small percentage of physician groups (11.21%) have an adequate number of patients for reliable measurement.

UPDATED TESTING

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

The results indicate that NQF 0555 is reliable at the health plan level, based on a sample of QHP products. Among the products with at least 30 denominator members, the average reliability was 0.70, which suggests sufficient signal relative to noise to discriminate performance between plans.

Reliability, Medicare PDPs

The results indicate that NQF 0555 is reliable at the health plan level, based on Medicare PDP data with at least 100 members in the denominator. In 2016, the majority of Medicare PDPs (83.6%) had at least 100 members in the denominator, which produced measure performance rates with sufficient reliability (0.70) to distinguish differences in performance among plans.

Based on the larger sample from the Medicare data, the reliability findings suggest that a denominator size of at least 100 members would be needed to achieve reliable results at the health plan level.

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*)

2b1.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

PRIOR SUBMISSION

Performance Measure Score

1. Convergent Validity - Relationship to another measure as expected (NQF 0555 compared to NQF 0556), Pearson Correlation Score
2. Systematic Assessment of Face Validity, Likert Scale, Overall Mean and Median Score (Discussed in 2.b2.3)
3. Threats to Validity, Analysis of Missing Data, Frequency

Convergent Validity: We compared a related NQF-endorsed measure, NQF 0556, which assesses INR monitoring after an interacting anti-infective drug is prescribed. We would expect the scores on these measures to be correlated since they reflect a similar concept of timely and appropriate INR monitoring. We tested the measure distributions for normality at each unit of analysis and then selected the appropriate statistical test for the distribution and assessed the significance of the correlation coefficient.

UPDATED TESTING

Convergent Validity: Using Pearson's correlation coefficients, we compared the performance of NQF 0555 with NQF 0541 (Proportion of Days Covered [PDC]: 3 Rates by Therapeutic Category), which has three rates of medication adherence and is part of the Medicare Part D Star Rating Program in 2015. NQF 0541 assesses adherence to medications for diabetes, hypertension, and cholesterol reduction (*Medication Adherence for Diabetes Medications*, *Medication Adherence for Hypertension [RAS Antagonists]*, and *Medication Adherence for Cholesterol [Statins]*). Our rationale for this comparison is as follows: plans with higher performance on medication adherence should have similar performance with INR testing, since both measures assess appropriate medication management.

PRIOR SUBMISSION

Face Validity Method: FMQAI's Technical Expert Panel (TEP) evaluated the face validity of the measure and measure score after field testing was completed. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®). TEP members were specifically asked whether "the performance score from the measure as specified represents an accurate reflection of quality of care." They responded by indicating their level of agreement with the statement on a 5-point Likert scale (1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree).

UPDATED TESTING

Face Validity: We systematically evaluated the face validity of NQF 0555 and the measure score after testing was completed. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®) with the Technical Expert Panel (TEP) advising the project. The TEP is composed of three representatives from large QHP issuers and nine representatives from other stakeholder groups, such as measurement industry representatives, clinical and nonclinical experts, and patient/caregiver representatives. TEP members were specifically asked whether they agree with the following statement: "The performance scores resulting from the measure NQF 0555 INR Monitoring for Individuals on Warfarin, as specified, can be used to distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy." They responded "yes" or "no," indicating either they did or they did not agree with the previous statement.

PRIOR SUBMISSION

Threats to Validity: Days' supply is a critical variable in determining warfarin usage. We assessed all warfarin claims for patients in the denominator for missing days' supply. Specifically, for missing days' supply, we analyzed the number (%) of beneficiaries in the measure denominator with one or more claims that had missing days' supply.

UPDATED TESTING

Threats to Validity: We examined the missingness of the prescription variable, *days' supply*, in our data.

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

PRIOR SUBMISSION

Convergent Validity: The measure rate is positively correlated with the NQF-endorsed measure, INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications (NQF 0556) at the ACO level (ρ=0.745, $p<0.0001$). The distribution of the measure rates is presented in Table 14.

Table 14. Distribution of Measure Rates – ACO

Measure	Count ACO	Mean Measure Rate	Standard Deviation	Median	Minimum	Maximum
INR Monitoring for Individuals on Warfarin (NQF 0555)	31	75.3%	9.8%	78.0%	58.5%	90.7%

INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications (NQF 0556)	31	21.7%	5.5%	21.4%	13.0%	32.9%

UPDATED TESTING

Convergent Validity: Results for NQF 0541 were available for 57 PDPs for the diabetes adherence rate and 58 PDPs for the hypertension and cholesterol medication adherence rates. The analysis revealed significant relationships between NQF 0555 measure scores and all three rates of medication adherence ($p<0.0001$ for all correlations; diabetes: $r=0.591$, hypertension: $r=0.700$, cholesterol: $r=0.751$). These results indicate positive linear associations with large effect sizes between NQF 0555 and three independent measure rates of medication adherence at the PDP level of analysis (Figures 1-3). According to Cohen's thresholds for product-moment correlations, 0.50 or higher is considered a large correlation.[1]

Citation:

1. Cohen J. A power primer. *Psychol Bull.* 1992;112(1):155-159.

Figure 1. Association Between Performance Rates for NQF 0555 and *Medication Adherence for Diabetes Medications*, Medicare PDPs, 2015

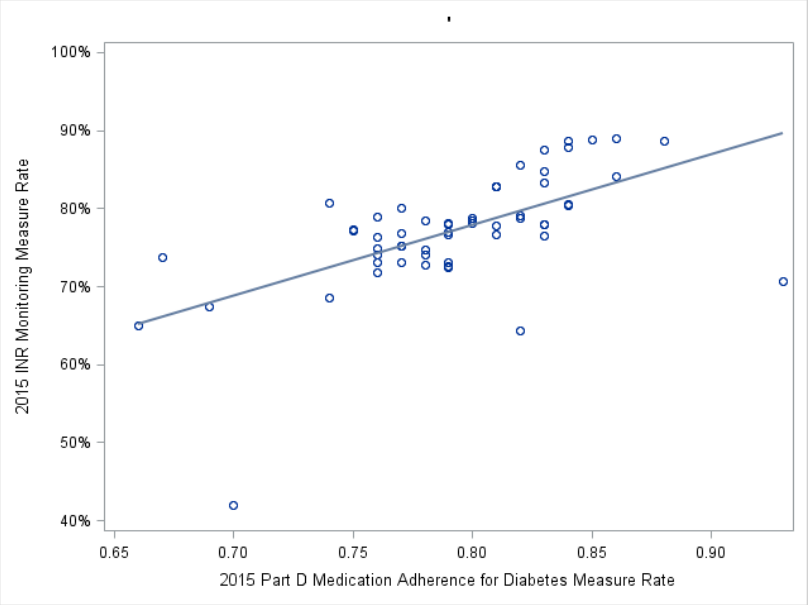


Figure 2. Association Between Performance Rates for NQF 0555 and *Medication Adherence for Hypertension (RAS Antagonists)*, Medicare PDPs, 2015

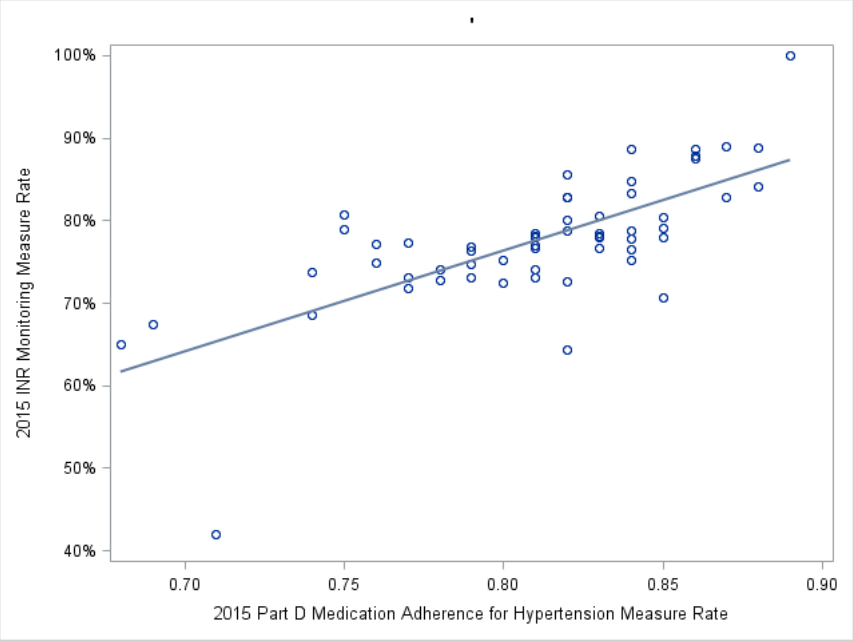
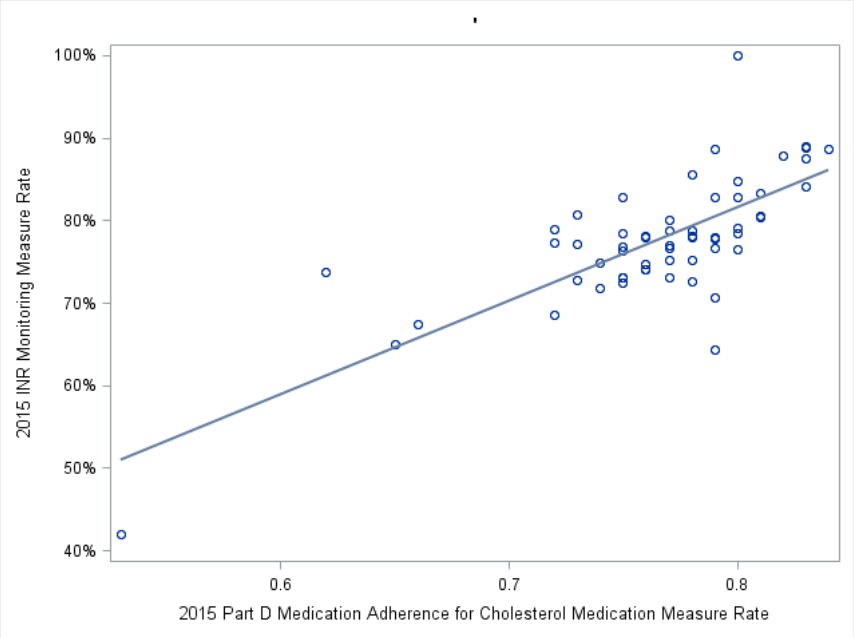


Figure 3. Association Between Performance Rates for NQF 0555 and *Medication Adherence for Cholesterol (Statins)*, Medicare PDPs, 2015



PRIOR SUBMISSION

Systematic Assessment of Face Validity: Fifteen of the 21 (71.4 %) TEP members completed the face validity evaluation for the measure. The results of the TEP rating of face validity on a scale of 1 to 5 are presented in Table 15.

Table 15. Results of the Face Validity Evaluation

Rating	Number of TEP (%)
5 (Strongly Agree)	4 (26.7%)
4 (Agree)	8 (53.3%)
3 (Neutral)	2 (13.3%)
2 (Disagree)	1 (6.7%)
1 (Strongly Disagree)	0

Of the TEP members who evaluated the measure for face validity, 80% (12/15) strongly agreed or agreed that the measure was valid as specified. The mean rate was 4, and the median rate was 4.

UPDATED TESTING

Systematic Assessment of Face Validity: Nine out of nine TEP members (100%) responding to the face validity survey agreed that NQF 0555 was valid as specified. Three TEP members did not complete the survey.

PRIOR SUBMISSION

Threats to Validity: Percentage of individuals in the denominator with one or more claims with missing days' supply - 0/263,080 (0%).

UPDATED TESTING

Threats to Validity: No individuals in either the QHP or Medicare PDP denominators had missing days' supply.

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?*)

PRIOR SUBMISSION

Convergent Validity: The measure rates between NQF 0555 and NQF 0556 were strongly correlated (>0.7) as expected, and this adds further support that the measure as specified is valid.

UPDATED TESTING

Convergent Validity: Performance comparison between NQF 0541, representing three rates of medication adherence, and NQF 0555 was strongly and positively correlated at the PDP level. The results support our hypothesized relationship between NQF 0555 and NQF 0541 and demonstrate that NQF 0555 is valid in capturing the quality of care related to medication management.

PRIOR SUBMISSION

Face Validity: In summary, 80% of TEP members who responded to the survey strongly agreed or agreed that the measure has face validity.

UPDATED TESTING

Face Validity: Of the TEP members who responded to the survey, 100% agreed that NQF 0555 can be used to distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy.

PRIOR SUBMISSION

Threats to Validity: All claims in the analysis had the days' supply field populated. Therefore, no impact on the accuracy of the measure is expected from missing days' supply.

UPDATED TESTING

Threats to Validity: Our evaluation of the days' supply field in both the QHP and Medicare PDP data resulted in zero missing values. Therefore, we conclude that missing data are not a threat to validity.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* ***2b4***

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*)

PRIOR SUBMISSION

Individuals with a home INR testing are excluded. To examine the effect of this exclusion, the measure rates with and without the exclusion were calculated and compared.

UPDATED TESTING

Individuals with home INR monitoring are excluded from the NQF 0555 denominator because not all of their INR tests are reliably captured in claims. The INR tests conducted at home are not submitted as individual claims. Therefore, the frequency of the INR tests cannot be ascertained for this population, which prohibits determining whether a home INR test was conducted within the 56-day timeframe specified by the numerator of this measure. To examine the effect of this exclusion, the measure rates with and without the exclusion were calculated and compared using data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs. QHP Issuer 4 did not have sufficient denominator sizes (n=30) for analyses and is thus not included in the results shown in Table 18.[1]

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

2015-2017 INR MONITORING AT HOME HCPCS CODES FOR EXCLUSION:
G0248 – Demonstrate Use Home INR Mon

G0249 – Provide Test Mats & Equip Home INR

G0250 – MD INR Test Review Inter Mgmt

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*)

PRIOR SUBMISSION

The exclusion was applied to the 10-state data from 2012. The aggregated denominator, numerator, and the measure rate across the 10 states are shown below in Table 16. In addition, Table 17 shows the results by states.

Table 16. Measure Rate by Exclusion Status

Home INR Excluded	Denominator	Numerator	Measure Rate	95% CI
Yes	263,080	193,606	73.6%	73.4%, 73.8%
No	281,812	196,757	69.8%	69.6%, 70.0%

Table 17. Exclusion Analysis by States

State	Excluding Patients with Home INR				Including Patients with Home INR			
	Den	Num	Rate	95% CI	Den	Num	Rate	95% CI
All	263,080	193,606	73.6%	73.4%, 73.8%	281,812	196,757	69.8%	69.6%, 70.0%
AZ	13,217	9,863	74.6%	73.9%, 75.4%	14,731	10,123	68.7%	68.0%, 69.5%
DE	4,028	3,039	75.5%	74.1%, 76.8%	4,371	3,091	70.7%	69.4%, 72.1%
FL	64,685	48,048	74.3%	73.9%, 74.6%	70,384	49,081	69.7%	69.4%, 70.1%
IA	23,399	19,466	83.2%	82.7%, 83.7%	23,979	19,554	81.6%	81.1%, 82.0%
IN	30,056	23,388	77.8%	77.3%, 78.3%	32,261	23,714	73.5%	73.0%, 74.0%
MO	27,245	20,787	76.3%	75.8%, 76.8%	29,290	21,093	72.0%	71.5%, 72.5%
MS	17,513	11,476	65.5%	64.8%, 66.2%	18,373	11,654	63.4%	62.7%, 64.1%
RI	3,828	3,392	88.6%	87.6%, 89.6%	4,051	3,445	85.0%	83.9%, 86.1%
TX	55,761	35,845	64.3%	63.9%, 64.7%	60,031	36,501	60.8%	60.4%, 61.2%
WA	23,348	18,302	78.4%	77.9%, 78.9%	24,341	18,501	76.0%	75.5%, 76.5%

For the overall cohort, the measure rate excluding patients with home INR is significantly higher than the measure rate including patients with home INR (95% confidence intervals do not overlap). For measure rates including and excluding home INR, there is a statistically significant difference between all pairwise comparisons of states ($p \leq 0.05$) except for Arizona, Delaware, Florida, and Missouri.

UPDATED TESTING

To determine the effect of the exclusion on the 2016 NQF 0555 measure rates, the rates were calculated with and without the exclusion, as shown in Table 18.

Table 18. 2016 INR Measure Rate by Exclusion Status

Product	Exclusion Status	Denominator	Numerator	Measure Rate	95% CI
QHP Issuer 1					
B	**No Exclusions**	328	143	43.6%	38.1%, 49.1%
B	**Home INR Monitoring Excluded**	326	143	43.9%	38.3%, 49.4%
QHP Issuer 2					
A	**No Exclusions**	205	122	59.5%	52.8%, 66.2%
A	**Home INR Monitoring Excluded**	203	120	59.1%	52.4%, 65.9%
QHP Issuer 3					
A	**No Exclusions**	185	105	56.8%	49.6%, 63.9%
A	**Home INR Monitoring Excluded**	185	105	56.8%	49.6%, 63.9%
B	**No Exclusions**	126	71	56.4%	47.7%, 65.0%
B	**Home INR Monitoring Excluded**	126	71	56.4%	47.7%, 65.0%
Medicare PDPs					
	No Exclusions	1,187,597	771,073	64.9%	64.8%, 65.0%
	Home INR Monitoring Excluded	1,059,826	754,993	71.2%	71.2%, 71.3%

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*)

PRIOR SUBMISSION

Statistically significant differences were identified in the measure rate with and without the exclusion of home INR monitoring. Since beneficiaries monitoring INR at home would not have claims for INR tests, this exclusion improves the measures validity.

UPDATED TESTING

Individuals with home INR monitoring are excluded from the NQF 0555 denominator because not all of their INR tests are reliably captured in the claims. The INR tests conducted at home are not submitted as individual claims. Furthermore, although two of the HCPCS codes used to identify home monitoring are for provision of INR test materials and physician review of test results, these two codes can be associated with up to four INR tests per claim. Therefore, the frequency of the INR tests cannot be accurately ascertained for this population. Our empirical analysis confirm that measure rates were lower in the Medicare population if the exclusion for beneficiaries was not applied because beneficiaries could meet the denominator definition of being on warfarin therapy for at least 56 days but did not meet the numerator since their monitoring of INR was conducted at home. For the QHP data, the rates did not differ significantly if the exclusion for members was not applied, because only a small number of individuals were conducting home INR monitoring. Therefore, we have

retained the measure exclusion, since patients form either population monitoring INR at home would not have reliable claims data for INR tests that could be used to satisfy the measure specifications (a test every 56 days).

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 2b5.

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

☒ No risk adjustment or stratification

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

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

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

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

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

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? Not applicable

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

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors? Not applicable

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.**

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

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

if stratified, skip to 2b4.9 Not applicable

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:** Not applicable

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): Not applicable

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves: Not applicable

2b3.9. Results of Risk Stratification Analysis: Not applicable

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*) Not applicable

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*)

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)*

PRIOR SUBMISSION

To identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the state, prescription drug plan, physician group, and ACO levels. Confidence intervals (CI 95%) were calculated around point estimates for each state, prescription drug plan, physician group, and ACO, and then compared to the overall mean of states, prescription drug plans, physician groups, and ACOs, respectively. If the confidence intervals did not overlap with the overall mean, the difference was considered statistically significant.

UPDATED TESTING

For this comprehensive re-evaluation, we used the same methods as described above for the evaluation of performance variation among QHP products and Medicare PDPs.

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*)

PRIOR SUBMISSION

Meaningful Differences at the State Level – 2012: Two of the 10 states (20.0%) had scores statistically significantly lower than the mean, and the other eight states (80.0%) had scores significantly higher than the mean. Measure rates ranged from 64.3% in Texas to 88.6% in Rhode Island, indicating suboptimal performance across all eight states (Table 19).

Table 19. 2012 State Level Performance

n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
10	75.8%	75.9%	64.3%	88.6%	7.2%	4.1%	64.9%	74.3%	75.9%	78.4%	85.9%

Meaningful Differences at the Plan Level – 2012: Of the plan scores, 33.3% of providers were statistically significantly lower than the mean, and 51.3% of providers were statistically significantly higher than the mean. For those plans with at least 100 eligible individuals, high- (90^{th} percentile) and low- (10^{th} percentile) performing plans were 18.9% apart, indicating suboptimal performance across plans and variations between high- and low-performing plans (Table 20).

Table 20. 2012 Prescription Drug Plan Level Performance

n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
39	74.5%	75.6%	59.7%	88.3%	7.2%	12.6%	64.8%	68.5%	75.6%	81.0%	83.6%

UPDATED TESTING

Meaningful Differences at the Plan Level – 2016

Measure rates across QHP products ranged from 43.9% to 59.1% (Table 21a) with a mean measure rate of 54.0%. These rates were substantially lower than the rate observed among Medicare PDPs (71.7%) (Table 21b).

Table 21a. 2016 QHP Performance for Those with at Least 30 Members in the Denominator

QHP Issuer	QHP Product	Rate	Confidence Interval
1	B	43.9%	38.3%, 49.4%
2	A	59.1%	52.4%, 65.9%
3	A	56.8%	49.6%, 63.9%
3	B	56.4%	47.7%, 65.0%

The reliability findings suggested that a denominator size of at least 100 members would be needed to achieve reliable results at the health plan level. Therefore, among Medicare PDPs with at least 100 denominator beneficiaries, we found that 41.2% (21/51) of plans had rates significantly lower than the mean, and 37.3% (19/51) of plans had rates significantly greater than the mean. For PDPs with at least 100 members, the difference in performance between high-performing (i.e., 90th percentile) and low-performing (i.e., 10th percentile) PDPs was 18.2%, indicating both variation between high- and low-performing PDPs and suboptimal performance across PDPs (Table 21b).

Table 21b. 2016 Medicare PDP Performance for Those with at Least 100 Members in the Denominator

n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
51	71.7%	71.4%	46.4%	85.1%	7.5%	10.1%	64.0%	67.3%	71.4%	77.4%	82.2%

PRIOR SUBMISSION

Meaningful Differences at the Physician Group Level – 2012: Of the physician group scores, 24.4% of providers were statistically significantly lower than the mean, and 28.1% of providers were statistically significantly higher than the mean, indicating a wide range of scores. For those physician groups with at least 50 eligible individuals, high- (90th percentile) and low- (10th percentile) performing physician groups were 28.4% apart. The results indicate ample room for improvement and meaningful differences in quality of care between the highest and lowest performing physician groups (Table 22).

Table 22. 2012 Physician Group Level Performance

n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
739	75.7%	77.2%	14.8%	97.1%	11.2%	15.4%	60.3%	68.9%	77.2%	84.2%	88.7%

Meaningful Differences at the ACO Level: Of the ACO scores, 41.9% of providers were statistically significantly lower than the mean, and 41.9% of providers were statistically significantly higher than the mean. For those ACOs with at least 50 eligible individuals, high- (90th percentile) and low- (10th percentile)

performing ACO were 24.5% apart, indicating suboptimal performance across ACOs and variation between high- and low-performing ACOs (Table 23).

Table 23. ACO Level Performance (2011)

n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
31	75.3%	78.0%	58.5%	90.7%	9.8%	18.5%	62.6%	65.5%	78.0%	84.0%	87.1%

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?*)

PRIOR SUBMISSION

The overall mean of ~75% of patients having an INR test every 56 days indicates that measure performance is suboptimal. Furthermore, across measurement units, there was ample variation in performance between high- and low-performing plans indicating room for improvement in INR monitoring rates.

UPDATED TESTING

The low performance rates of the QHP products (average rate of 54.0% in 2016) suggests substantial opportunity for improvement in the management of patients on warfarin among QHPs in the Health Insurance Exchanges. Among Medicare PDPs, measure rates decreased from 2012 to 2016. In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 71.7%) among Medicare PDPs. The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries on warfarin therapy.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*) Not applicable

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

Not applicable

2b5.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*) Not applicable

---

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

UPDATED TESTING

NQF 0555 is a claims-based measure and relies on final paid claims from payors (Medicare, QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, or QHP Issuer 4). The most critical data element that could lead to missing cases, *days' supply of medication*, was complete in the datasets used for testing. None of the claims contained missing data for the element *days' supply of medication*.

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*)

Not applicable

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

Not applicable

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

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).

This measure is specified using administrative claims data. At this time, there is no plan to specify the measure as an eCQM.

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

Attachment:

3c. Data Collection Strategy

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

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

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

Testing demonstrated the measure was feasible to be specified and calculated using administrative claims data from QHP products and Medicare PDPs. Data used in the calculation of this measure are obtained from administrative claims, which are routinely, reliably, and securely collected for billing purposes. We do not anticipate any feasibility or implementation issues related to data collection for this measure. No threats to the validity of this measure were identified using a limited analysis designed to address missing data.

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 (Internal to the specific organization)	

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

Not applicable because the measure is not currently in use.

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?*)

This measure was previously in use for the Quality and Resource Use Reports,[1] but has not been in use since the last NQF review in 2013. This measure is now being considered for use in the Quality Rating System for QHPs. The Quality Rating System is intended to inform consumers when choosing a QHP from the Health Insurance Exchange by providing comparisons of the quality of care provided by each health plan. The Quality Rating System is not used for payment or penalty to the health plans.

Citations

1. Centers for Medicare & Medicaid Services. Analysis of 2011 Physician Feedback Program Individual Reports. Retrieved July 19, 2018, from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/PY2011-Individual-Report.pdf.

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.*)

This measure will be considered for use in the Quality Rating System for QHPs offered on the Exchanges.

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 measure is not currently implemented in a public reporting program, and therefore there is no information available regarding feedback during implementation. Our Technical Expert Panel (TEP) reviewed the updated measure evidence, testing, and performance results and interpretation via several webinar conferences. The TEP is comprised of three representatives from large QHP issuers, and nine individuals from other stakeholder groups, such as organization representatives, clinical and nonclinical experts, and patient/caregiver representatives. A full list of the TEP members' names and organizations is noted under the Additional Information section of this document.

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.

Meetings with the TEP were held throughout 2015-2017. During these meetings, TEP members were provided with updated measure evidence, development, and testing results. Data necessary to judge the validity and usability of the measure were provided, along with the measure algorithm and a complete list of codes used to calculate the measure. Questions that arose from these meetings were addressed either during the meeting or in follow-up communications.

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.

TEP members were encouraged to provide feedback throughout the measure re-evaluation process by means of meeting discussions and voting and through follow-up communications. Members were sent a questionnaire focused on face validity and usability which contained closed-ended response options and free text comment fields.

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

TEP members who represented organizations being measured responded to a questionnaire and indicated that this measure has clear and precise specifications, that health plans will use information from this measure to improve quality of care for patients on warfarin therapy, and that health plans will be able to implement the measure without undue burden for reporting for the Quality Rating System. Those being measured agreed (n=3/3) that the measure can distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy (i.e., the measure has face validity). Additionally, one issuer compared this measure against a warfarin measure they currently use (to measure INR re-check intervals) and found this measure, INR Monitoring for Individuals on Warfarin, to be valid when compared to their current measure.

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

TEP members who represented other stakeholder groups responded to a questionnaire and indicated that this measure has clear and precise specifications, that health plans will use information from this measure to improve quality of care for patients on warfarin therapy, and that health plans will be able to implement the measure without undue burden for reporting for the Quality Rating System. Respondents agreed (n=6/6) that the measure can distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy (i.e., the measure has face validity). Three members did not respond to the questionnaire. All feedback received regarding this measure indicates that the measure will be useful for health plans to improve quality of care for patients on warfarin therapy and can be implemented without undue burden for reporting for the Quality Rating System.

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.

TEP and workgroup feedback was considered throughout the measure re-evaluation process. Feedback received was unanimously (9/9) in favor of the specifications described in this submission form; therefore, revision of the measure specifications was not necessary.

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.

The measure is not currently implemented in a public reporting program; therefore, we describe how the performance results could be used to further the goal of high-quality care. The low performance rates of the QHP products in our analysis (average rate of 54.0% in 2016) suggest substantial opportunity for improvement in the management of patients on warfarin among QHP products in the Health Insurance Exchange. Among Medicare PDPs, measure rates decreased from 74.5% in 2012 to 71.7% 2016, underscoring the need for performance measurement for patients on warfarin therapy. The performance rates of this measure in both populations suggest opportunity for improving care for patients on warfarin therapy.

This measure is actionable by both providers and plans and can be used to further the goal of high-quality care. The desired outcome for this measure is fewer bleeding and thromboembolic events in individuals on warfarin. Regular INR monitoring is associated with increased time in therapeutic range [1-3] and reduced risk of thromboembolism,[3] whereas subtherapeutic INR is correlated with significantly higher total healthcare

costs[4, 5] and greater risks of stroke/SE,[6] major bleeding[6,7], thromboembolism,[7] and mortality.[6-8] Health outcome linkage is further discussed in the Evidence Attachment.

Citations

1. Rose AJ, Miller DR, Ozonoff A, et al. Gaps in monitoring during oral anticoagulation: insights into care transitions, monitoring barriers, and medication nonadherence. Chest. 2013;143(3):751-757. doi: 10.1378/chest.12-1119.

2. Rose AJ, Park A, Gillespie C, et al. Results of a regional effort to improve warfarin management. Annals of Pharmacotherapy. 2017. doi: 10.1177/1060028016681030.

3. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. Thromb Res. 2013;132(2):e124-130. doi: 10.1016/j.thromres.2013.06.006.

4. Nelson WW, Wang L, Baser O, Damaraju CV, Schein JR. Out-of-range international normalized ratio values and healthcare cost among new warfarin patients with non-valvular atrial fibrillation. Journal of medical economics. 2015;18(5):333-340. doi: 10.3111/13696998.2014.1001851.

5. Deitelzweig S, Evans M, Hillson E, et al. Warfarin time in therapeutic range and its impact on healthcare resource utilization and costs among patients with nonvalvular atrial fibrillation. Curr Med Res Opin. 2016;32(1):87-94. doi: 10.1185/03007995.2015.1103217.

6. Liu S, Li X, Shi Q, et al. Outcomes associated with warfarin time in therapeutic range among US veterans with nonvalvular atrial fibrillation. Curr Med Res Opin. 2018;34(3):415-421. doi: 10.1080/03007995.2017.1384370.

7. Labaf A, Sjalander A, Stagmo M, Svensson PJ. INR variability and outcomes in patients with mechanical heart valve prosthesis. Thromb Res. 2015;136(6):1211-1215. doi: 10.1016/j.thromres.2015.10.044.

8. Schein JR, White CM, Nelson WW, Kluger J, Mearns ES, Coleman CI. Vitamin K antagonist use: evidence of the difficulty of achieving and maintaining target INR range and subsequent consequences. Thromb J. 2016;14:14. doi: 10.1186/s12959-016-0088-y.

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.

This measure is not currently in use, however, previous measure maintenance efforts reported that no unintended negative consequences had been identified in the 2011 Quality and Resource Use Reports.[1]

Citations

1. Centers for Medicare & Medicaid Services. Analysis of 2011 Physician Feedback Program Individual Reports. Baltimore, MD: Centers for Medicare & Medicaid Services; 2012. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/PY2011-Individual-Report.pdf. Accessed September 12, 2018.

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

The measure has not been in use, and therefore there are no unexpected benefits from implementation to report.

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)

0556 : INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications

2732 : INR Monitoring for Individuals on Warfarin after Hospital Discharge

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

Related measures are endorsed.

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?

Yes

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

The measure under review (NQF 0555) is related to both NQF 0556 (INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications) and NQF 2732 (INR Monitoring for Individuals on Warfarin after Hospital Discharge). All three have the same measure focus, which is INR testing, and their specifications for INR testing are harmonized; however, the three measures have different clinical foci and target populations. The measure under review (NQF 0555) focuses on INR testing during every 56-day interval in which an individual is prescribed warfarin. NQF 0556 focuses on INR testing within three to seven days for patients on warfarin who are prescribed anti-infective medications that are known to interact with warfarin and result in a higher risk for adverse events, and NQF 2732 focuses on INR monitoring within 14 days of hospital discharge for individuals on warfarin who were not yet in the therapeutic range at the time of discharge. Due to the difference in the clinical foci, the timeframe for INR monitoring (three to seven days, 14 days, 56 days) is different among the three measures and complimentary rather than competing with one another.

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.)

Not applicable

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.

No appendix **Attachment:**

Contact Information

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

Co.2 Point of Contact: Sophia, Chan, Sophia.Chan@cms.hhs.gov

Co.3 Measure Developer if different from Measure Steward: Health Services Advisory Group

Co.4 Point of Contact: Melissa, Castora-Binkley, mcastora-binkley@hsag.com, 813-865-3182-

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.

Original Technical Expert Panel (TEP), 2009-2011

1. Douglas Bell, MD, PhD, Associate Professor in Residence, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Researc
2. Jill S. Borchert, PharmD, BCPS, FCCP, Professor, Pharmacy Practice & PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy?
3. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association
4. Jannet Carmichael, PharmD, BCPS, FCCP, FAPHA, VISN 21 Pharmacy Executive, VA Sierra Pacific Network
5. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago
6. Edward Eisenberg, MD, Vice President and Chief Medical Officer, Medicare, Medco Health Solutions
7. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.
8. David Nau, PhD, MS, Senior Director of Research & Performance Measurement, PQA, Inc.
9. N. Lee Rucker, PhD, MS, Senior Strategic Policy Advisor, AARP - Public Policy Institute
10. Marissa Schlaifer, RPh, MS, Director of Pharmacy Affairs Academy of Managed Care Pharmacy
11. Brad Tice, PharmD, Chief Clinical Officer, PharmMD Solutions, LLC
12. Jennifer K. Thomas, PharmD, Manager, Pharmacy Services, Delmarva Foundation for Medical Care/Delmarva Foundation of the District of Columbia
13. Darren Triller, PharmD, Director, Pharmacy Services, IPRO
14. Neil Wenger, MD, MPH, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research

Current Technical Expert Panel (TEP), 2015-2017

1. Andy Amster, MSPH, Kaiser Permanente National Office
2. Marybeth Farquhar, PhD, MSN, RN, URAC
3. Susan Fitzpatrick, RN, BSN, Cigna Healthcare
4. Aparna Higgins, Duke-Margolis Center for Health Policy; Brandeis University
5. Jon Mark Hirshon, MD, PhD, MPH, University of Maryland, School of Medicine
6. Christine Hunter, MD, US Office of Personnel Management
7. Carol Keegan, PhD, Patient representative
8. Dana Mukamel, PhD, University of California, Irvine
9. Chinwe Nwosu, NS, America's Health Insurance Plans
10. Derek Robinson, MD, MBA, FACEP, Health Care Service Corporation
11. Arlene Salamendra, Patient representative
12. Ted von Glahn, MSPH, von Glahn Consulting

The TEP evaluated this medication safety measure drafted by Health Services Advisory Group (HSAG), originally developed by FMQAI, in regard to the four primary measure evaluation criteria used in the NQF consensus endorsement process (importance, scientific acceptability, feasibility, and usability). The TEP discussed the strengths and weaknesses of the measure and made recommendations regarding measure specifications, and inclusion and exclusion criteria.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2009

Ad.3 Month and Year of most recent revision: 07, 2018

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

Ad.5 When is the next scheduled review/update for this measure? 07, 2019

Ad.6 Copyright statement: Limited proprietary coding is contained in the Measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets.

Ad.7 Disclaimers: This performance measure does not establish a standard of medical care and has not been tested for all potential applications.

Ad.8 Additional Information/Comments: Not applicable

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

Brief Measure Information

NQF #: 0753

Measure Title: American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure

Measure Steward: Centers for Disease Control and Prevention

Brief Description of Measure: Facility adjusted Standardized Infection Ratio (SIR) and Adjusted Ranking Metric (ARM) for deep incisional and organ/space Surgical Site Infections (SSI) at the primary incision site among adult patients aged >= 18 years as reported through the CDC National Health and Safety Network (NHSN).

Developer Rationale: It is envisioned the use of this measure will promote SSI prevention activities which will lead to improved patient outcomes including reduction of avoidable medical costs, and patient morbidity and mortality.

Numerator Statement: Deep incisional primary (DIP) and organ/space SSIs during the 30-day postoperative period among patients = 18 years of age, who undergo inpatient colon surgeries or abdominal hysterectomies. SSIs will be identified before discharge from the hospital, upon readmission to the same hospital, or during outpatient care or admission to another hospital (post-discharge surveillance).

Numerator Exclusion SSI events with PATOS* field = yes.

Infection present at time of surgery (PATOS): PATOS denotes that there is evidence of an infection or abscess at the start of or during the index surgical procedure (in other words, it is present preoperatively). PATOS is a YES/NO field on the SSI Event form. PATOS does not apply if there is a period of wellness between the time of a preoperative condition and surgery. The evidence of infection or abscess must be noted/documented intraoperatively in an operative note or report of surgery. Only select PATOS = YES if it applies to the depth of SSI that is being attributed to the procedures (e.g., if a patient has evidence of an intraabdominal infection at the time of surgery and then later returns with an organ/space SSI the PATOS field would be selected as a YES. If the patient returned with a superficial or deep incisional SSI the PATOS field would be selected as a NO). The patient does not have to meet the NHSN definition of an SSI at the time of the primary procedure but there must be notation that there is evidence of an infection or abscess present at the time of surgery. PATOS is not necessarily diagnosis driven.

Denominator Statement: An NHSN Operative Procedure is a procedure:

• that is included in the ICD-10-PCS or CPT NHSN operative procedure code mapping. And

• takes place during an operation where at least one incision (including laparoscopic approach and cranial Burr holes) is made through the skin or mucous membrane, or reoperation via an incision that was left open during a prior operative procedure And

• takes place in an operating room (OR), defined as a patient care area that met the Facilities Guidelines Institute's (FGI) or American Institute of Architects' (AIA) criteria for an operating room when it was constructed or renovated. This may include an operating room, C-section room, interventional radiology room, or a cardiac catheterization lab.

Exclusions: Otherwise eligible procedures that are assigned an ASA score of 6 are not eligible for NHSN SSI surveillance.

Using multivariable logistic regression models for colon surgeries and abdominal hysterectomies, the predicted number of SSIs is obtained. These predicted numbers are summed by facility and surgical procedure and used as the denominator of this measure (see also 2a.8).

Denominator Exclusions: Denominator data are excluded from the SSI measure due to various reasons related to data quality, data outlier and data errors. The complete list of universal exclusion criteria applied to denominator are listed in the SSI section of the SIR guide that is referenced above. These exclusions include but are not limited to procedures associated with SSI events where the PATOS = yes, and those with ASA Class VI (6). The measure specific denominator exclusions for the Complex 30-day SSI, are off plan colon and abdominal hysterectomy procedures, procedures performed on persons under the age of 18, and procedure performed on an outpatient basis. .

Note: Under the 2015 baseline, both primarily closed procedures and those that are not closed primarily are included in the denominator data.Persons under the age of 18, those having a procedure performed on an outpatient basis, procedures associated with SSI events where the PATOS = yes, those with ASA Class VI (6) are excluded.

Note: Both primarily closed procedures and those that are not closed primarily are included in the denominator data.

Measure Type: Outcome

Data Source: Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

Level of Analysis: Facility, Other, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2012 **Most Recent Endorsement Date:** Jan 17, 2012

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 provided a summary of the link between preoperative and postoperative interventions with fewer incidence of surgical site infections. The developer envisions the use of this measure will promote SSI prevention activities which will lead to improved patient outcomes including reduction of avoidable medical costs, and patient morbidity and mortality.
- The developer provided the following updated 2017 clinical guidelines
    - Berríos-Torres, SI. et al., Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection. JAMA Surg, 152(8): (2017):784-791
        - Based on 170 studies, this 2017 Guideline provided recommendations for preventing surgical site infections including but not limited to use of sterile technique, avoidance of preoperative shaving of the operative site, preoperative decontamination of the surgical site, administration of preoperative prophylactic antibiotics within a prescribed timeframe, maintaining glycemic control in diabetic patients, and providing an increased inspired fraction of oxygen to the patient during and immediately following surgery

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:

The developer provided The Center for Disease Control and Prevention's Guideline for the Prevention of Surgical Site Infection, which has been published since the last NQF review.

- Berríos-Torres, SI. et al., Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection. JAMA Surg, 152(8): (2017):784-791.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

Outcome measure (Box 1) → Empirical data shows relationship between outcome and one healthcare action provided (Box 2) → Empirical data passes for clinical evidence.

RATIONALE:

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer provided performance gap data in SIRs across facilities and nationally for both Adominal Hysterectomy Surgical Site Infections and Colon Surgery Surgical Site Infections for 2015 and 2016

- Abdominal Hysterectomy Surgical Site Infections (Facilities and National):
    - Facilities
    - 2015 Facility SIRs range-0.00 – 2.710 (median: 0.762)
        - 2016 Facility SIRs range-0.00 – 2.513 (median: 0.722)
    - National
        - National SSI HYST SIR in 2015 is 0.989 = 2,432 observed / 2,459.654 predicted SSIs
        - National SSI HYST SIR in 2016 is 0.868 = 2,138 observed / 2,462.289 predicted SSIs
        - Percent Change: 2016 v. 2015 12% decrease
- Colon Surgery Surgical Site Infections (Facilities and National):
    - Facilities
        - 2015 Facility SIRS range-0.00 – 2.399 (median: 0.783).
        - 2016 Facility SIRS range-0.00 – 2.170 (median: 0.781).
    - National
        - National SSI COLO SIR in 2015 is 0.989 = 8010 observed / 8,102.668 predicted SSIs
        - National SSI COLO SIR in 2016 is 0.931 = 7,960 observed / 8,553.309 predicted SSIs
        - Percent Change: 2016 v. 2015 6% decrease
- In addition, the developer cited literature/reports that display the status of Healthcare Associated Infections (HAIs) in the United States over time and currently.
    - The Healthcare-associated Infections in the United States, 2006-2016: A Story of Progress located here: https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html
    - The 2015 National and State Healthcare-associated Infection Data Report: https://www.cdc.gov/hai/surveillance/data-reports/2015-HAI-data-report.html

Disparities

- Per developer, there are no studies providing evidence of a direct relationship between social risk and HAIs. However per developer, currently the evidence is not strong enough to adjust measure for social risk factors.
- However the developer does note some evidence to the contrary is available.
    - For example, empirical data analyzed by CDC, specifically surgical site infection (SSI) data that hospitals report to NHSN, have yielded findings that SSI risk is lower among older patients compared to younger patients for some surgical procedures (http://www.cdc.gov/nhsn/pdfs/pscmanual/ssi_modelpaper.pdf)
    - The developer cited literature that certain patient-related factors have been associated with an increased risk of SSI, using the Complex 30-day SSI model,
        - For Colon Surgery: advanced age, American Society of Anesthesiologists' physical status classification (ASA) >2, diabetes, gender, obesity, procedure closure technique and procedures performed at oncology facilities vs. those performed at non oncology facilities.
        - For Abdominal Hysterectomy: younger age, increasing ASA classification, obesity, diabetes, and procedures performed at oncology facilities vs. those performed at non-oncology facilities.

Questions for the Committee:

- Specific questions on information provided for gap in care.

- Is there a gap in care that warrants a national performance measure?
- Since no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?

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

RATIONALE:

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

1a. Evidence
Comments:

**Data provided apply directly to the measure. The measure relates to the desired outcome of reducing/eliminating post-surgical site infections for the specific procedures included.

**It is.

**Pass, an outcomes metric, clinical practice guidelines.

**Hysterectomy seems to have stronger evidence; colons not as strong.

**Outcome measure with empirical data passing for clinical evidence.

**High.

**Acceptable.

**evidence is acceptable and updated evidence provided.

**Patient Reported Outcome Measure. Originally endorsed 2012. An update of The Center for Disease Control and Prevention's Guideline for the Prevention of Surgical Site Infection has been published since the last submission.

1b. Performance Gap
Comments:

**A performance gap continues to exist. No disparities data was offered.

**Yes.

**Moderate, two year data provided comparing facilities and national trends, to revisit health disparities when more studies are available establishing direct link between social risk and HAIs.

**Performance gap higher for colons-no studies on disparities.

**Performance gap exists.

**High.

**Appears there are some really poor performing facilities.

**Performance gap noted.

**When SIRs are compared over time, assessment of performance can be made. CDC has demonstrated significant performance gaps in SIRs across facilities. Evidence of gaps were discussed. • The developer provided performance gap data in SIRs across facilities and nationally for both Adominal Hysterectomy Surgical Site Infections and Colon Surgery Surgical Site Infections for 2015 and 2016. • In addition, the developer cited literature/reports that display the status of Healthcare Associated Infections (HAIs) in the United States over time and currently. However per developer, currently the evidence is not strong enough to adjust measure for social risk factors.

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.

Composite measures only:

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Method Panel Review (Combined)

Evaluation of Reliability and Validity (and composite construction, if applicable):

Scientific Methods Panel Votes: Consensus Not Reached on Reliability

- Reliability: H-0, M-2, L-1, I-2
- Validity: H-1, M-4, L-0, I-0

Reliability

- Testing was conducted at the data element and measure score levels.
- Limited element **validity** testing was conducted for the population (state) level of analysis. If accepted as adequate data element validation, no additional reliability testing for the population (state) level of analysis is required**.**
- Measure Score (for facility level of analysis)
    - The developer stated that "*Reliability was estimated as the between- facility variance from a generalized linear mixed model divided by the total variance estimated from the same model*." The Methods Panel would have liked more detail about the methodology used in testing.
        - Results were colorectal surgeries were incomplete (which disturbed panel members) and mean reliability estimates were fairly low:

        - For COLO SSI, mean reliability=50.1%; xxx of 2,009 facilities met the Minimum Precision Criteria (MPC) and had reliability exceeding 40%. Developer noted that "Around one-third of facilities that met the MPC had reliability below the commonly-used 40% threshold for COLO SSI".
        - For HYST SSI, mean reliability=52.9%; and 652 of 787 facilities meeting the MPC had reliability exceeding 40%.
- Although not required by NQF at this time, Methods Panel members also noted that an analysis of the "stability" of the measure results for the facility level of analysis (i.e., that facility scores and rankings would not change dramatically in the short term) would have been helpful due to the rarity of the outcomes being measured.

Validity

- This measure is risk adjusted and a limited amount of data element validity testing was conducted. Although the panel rated validity as "Moderate", members had several concerns.
- Methods Panel members voiced concern about exclusions to the denominator that are due to "data quality, data outliers, and data errors". Members were concerned with the assertion that "*missing data is not a problem*" if incomplete reporting by facilities (which may be related to quality) means that they are not being assessed by the measure.
- Many facilities did not meet the Minimum Precision Criteria (MPC), which requires a facility to have at least one predicted event from the risk-adjustment model. This underscores the rarity of the outcomes being assessed.
- Risk model calibration (2b3.7) for SSI HYST shows a potential issue with Hosmer Lemeshow test (p = 0.012). Panel members noted that decile plots could help illuminate whether there is a problem with model calibration.
- Data element testing (for facility and population level of analysis)
    - Conducted limited data element validity testing for 7 states for the COLO SSI outcome and for 3 states for the HYST SSI outcome. It was not completely clear what element(s) were tested (likely the numerator only).
    - Panel members also noted that the relatively low sensitivity results, noting that the submission indicates there are disincentives for reporting colon surgery SSIs to the NHSN
        - Results: COLO – Mean measurements identified (ranges) were as follows:
            - Sensitivity: 74.9% (59.8-90.1)
            - Specificity: 99.1 % (98.7-100)
            - Positive Predictive Value: 95.8% (91.7-100)
            - Negative Predictive Value: 93.5% (85.3-97.2)
        - Results: HYST – Mean measurements identified (ranges) were as follows:
            - Sensitivity: 80.7% (75.4-100)
            - Specificity: 98.9% (88.9-99.1)
            - Positive Predictive Value: 92.6% (91.5-94.4)
            - Negative Predictive Value: 96.9% (96.9-100)
- Developers did not conduct data element validation for the variables used in the risk-adjustment model. Panel members think these are critical data elements that should be tested.
- At least one panel member thought that the risk-adjustment approach might not be robust enough, but believes the Standing Committee should weigh in from a clinical perspective.
- Another panel member expressed frustration about lack of reporting of the "optimism statistic", which apparently would provide information about the utility of the risk-adjustment approach (the submission indicates this analysis was conducted but the results were not provided).

Additional Information Provided by Developer on 12/17/18 in Response to the Scientific Methods Panel Evaluation above:

- Reliability was estimated as the between- facility variance from a generalized linear mixed model divided by the total variance estimated from the same model. For COLO SSI, the mean reliability for facilities meeting the Minimum Precision Criteria (MPC) of at least one predicted event was 50.1% and 1,323 of 2,009 (66%) facilities had reliability exceeding 40%. For HYST SSI, the mean reliability for facilities meeting the MPC was 52.9% and 652 of 787 (83%) facilities meeting the MPC had reliability exceeding 40%.
- The optimism statistics from the validation for the Colon (COLO) and abdominal hysterectomy (HYST) Complex 30-day SSI measures are 0.0006 and 0.00145, respectively.
- Reliability testing of SSI COLON and HYST procedures are estimated from the chart review studies conducted by the state health departments. These validation studies are primarily focused on numerator validations of SSI events and few states have also conducted validation of risk factors. While NHSN provides a guidance toolkit for SSI validations for external agencies, due to resource constraints these above mentioned studies vary in validation methodology. Hence it is not feasible to extrapolate the reasons for varying estimates of reported accuracies to specific generalizable reasons.

Standing Committee Action Item(s):

- The Standing Committee must discuss the reliability testing and determine if the results are acceptable.
- The Standing Committee should discuss validity, but could agree to accept the ratings of the Scientific Methods Panel. The Standing Committee also could recommend that developers conduct additional analyses recommended by the Panel, perhaps by time of the next maintenance evaluation.

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 did not reach consensus with the reliability testing for the measure. The Committee will need to discuss and 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 did not reach consensus with the testing for the measure. The Committee will need to discuss and vote on reliability.

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient** ☒ **Consensus Not Reached by Scientific Methods Panel**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient** ☐ **Consensus Not Reached by Scientific Methods Panel**

Combined Methods Panel Evaluation: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

*Note: Completed by multiple Scientific Methods Panel members and therefore multiple responses provided in checkboxes.

Measure Number: 0753

Measure Title: American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

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

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☒ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

   PANEL MEMBER 1: The specs are quite precise with the numerator being the number of specifically defined surgical site infections in a defined population who undergo defined procedures and the denominator is predicted number of such events. A ratio of the two is calculated.

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

   PANEL MEMBER 1: I have no concerns

   PANEL MEMBER 2: Exclusions in denominator include "data quality, data outlier and data errors". Is poor recording not a possible indicator of poor quality?

   PANEL MEMBER 3: Sensitivity generally high but varies, specificity, PPV, NPV generally high.

   PANEL MEMBER 4: The measure specifications are very detailed. My one struggle is the acronyms, such as ASA and PATOS. This makes it hard to always follow which cases are excluded/included in this measure.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**

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

Submission document: Testing attachment, section 2a2.2

PANEL MEMBER 1: It is mentioned that the SSI data used in this measure is the same as the SSI data utilized in a previously NQF endorsed measure and is a widely accepted method for summarizing mortality experience.
Nonetheless, between-facility variance was estimated from a generalized linear mixed model divided by the total variance estimated from the same model. For one measure, the COLO SSI, the mean reliability for facilities meeting the Minimum Precision Criteria of at least one predicted event was 50.1%. One-third of facilities were below the 40% level. For HYST SSI, the mean reliability was 52.9%. The data is summarized but not methodology is not mentioned.

PANEL MEMBER 2: Reliability estimates are given as the ratio of between facility variance to total variance explained based on generalized linear mixed models for colorectal surgeries and hysterectomies, a reasonable approach. It would have been valuable to understand what other variables were included and their contribution to the explained variance, as well as the total variance actually explained.

PANEL MEMBER 3:

The basic construction of the measure is:.

Get count SSI for COLO or HYST

Estimate expected count from logistic regression.

Construct actual to expected by dividing actual rate to estimated rate.

Potential sources of unreliability of measure:

a. Errors in counts of events.
b. Variability in counts over time due to random fluctuation
c. Imprecision in risk adjustment model

Methods used:

a. Reabstracting and assessment of sample of charts and calculation of specificity, sensitivity, positive predictive value, negative predictive value. Method is appropriate.
b. Not done. Do not provide range of counts, so cannot estimate likely year to year variability due to randomness.
c. Variables in regression model initially identified through expert panels based on available data taking burden into account. Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but values not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on "optimism in the fit." No assessment made of stability of SIRs across bootstraps or change in ranking across significantly higher, significantly lower, although only 5% in higher and 6% in lower category for COLO where enough cases to categorize, and for HYST, 4% higher, 4% lower.

   c-stats from risk adjustment models low (<.6), but that suggests risk adjustment explaining small amount of variance. However, given this, failure to consider how ranking changes given sizable variation in patient level predicted probabilities in 95% CI of risk model (3 times for COLO, 5 times for HYST) makes assessment of reliability limited.

   Homer-Lemeshow p value 0.012 for HYST, suggesting problems of goodness of fit.

PANEL MEMBER 4: The methods are from the original submission and seem appropriate for the measure.

PANEL MEMBER 5: Between/total variability in GLM – ICC

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

Submission document: Testing attachment, section 2a2.3

PANEL MEMBER 1: As mentioned above, reliability is about 50% for two measures and 1/3 of the facilities are below 40% for one of the measures.

PANEL MEMBER 2: Section 2a2.3 is a bit confusing. The reliability for facilities exceeding their 40% threshold states "x of 2009 facilities" for colorectal surgeries. For hysterectomies the reliability data are given as means and proportions for those "meeting the minimum precision criteria", which appears to be only 787 of 3250 facilities (24.2%). So the rest do not have at least one "predicted event"? How is this interpreted? Volume-outcome issues?

PANEL MEMBER 3: Sensitivity, specificity, PPV, NPV okay. Failure to consider year to year variability due to randomness is a weakness.

The developers do not report the "optimism" estimate or any other data to assess the stability of the SIRs or rankings over the bootstrapped models. The 95% CIs are generally narrow, but still result in a 3 fold difference in estimated risk across CI range for COLO and 5 fold for HYST.

PANEL MEMBER 4: There seems to be an issue with facility size/number of cases and reliability. This could reflect underlying data collection sophistication at different facilities. This is also likely to change over time, so new analysis of reliability would be helpful.

PANEL MEMBER 5: Appropriate, but levels at the low end of acceptability (50%)

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 <u>all</u> testing results):

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

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

☐ **Low** (NOTE: Should rate 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: The submission provided the summary data for two outcomes but alluded to the methodology as a between-facility variance from a generalized linear mixed model divided by the total variance estimated from the same model. With this methodology the results obtained were noted in Q7.

PANEL MEMBER 2: See #7.

PANEL MEMBER 3: There is inadequate information presented on the stability of the measure to reasonable expected variation in counts of SSIs or the precision and stability of the risk adjustment model.

PANEL MEMBER 4: Although the data is not presented, the correlation between facility size and reliability raises concerns. For the colon measure in particular, two-thirds of sites have reliability below their own target (40%), suggesting moderate reliability.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

PANEL MEMBER 1: I have no concerns. Infection present at the time of surgery accounted for 2.15% of the COLO patients and 1.37% of the HYST patients.

PANEL MEMBER 2: See specifications (#2 above).

PANEL MEMBER 3: None

PANEL MEMBER 4: It seems somewhat complicated to determine which surgical events count towards the numerator in this measure, but that is squarely a clinical question.

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

Submission document: Testing attachment, section 2b4.

PANEL MEMBER 1: It is commented, “we see variations”, and “we can identify facilities for which summary measure warrants investigation and response”. For COLO patients 2% of the total, 5% with SIR, had a ratio greater than 1, while 3% of the total, 6% with SIR, were less than 1. For HYST, the numbers were 1 and 4%, and 1 and 4%, respectively.

PANEL MEMBER 2: There are relatively large discrepancies between rates for those with and without SIR: - how are the data used?

PANEL MEMBER 3: See discussion re reliability in items 7a,c, 11.

PANEL MEMBER 4: The authors show cross-sectional differences between facilities and improvement in the measures over time – seems like good evidence of meaningful difference. That said, the rates are quite low. At some point the measure will bottom out.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

Submission document: Testing attachment, section 2b5.

PANEL MEMBER 1: The data source is standardized and validated.

PANEL MEMBER 3: NA

PANEL MEMBER 4: NA

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

Submission document: Testing attachment, section 2b6.

PANEL MEMBER 1: I have no concerns with missing data. Facilities are prevented from entering incomplete records.

PANEL MEMBER 2: See #29.

PANEL MEMBER 3: NA

PANEL MEMBER 4: None.

16. **Risk Adjustment**

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

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

☒ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

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

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

PANEL MEMBER 4: - its complex.

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☒ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No

PANEL MEMBER 4: – surgical technique

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☒ No

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

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

PANEL MEMBER 1: An expert panel was formed to identify potential risk factors in the initial phase of model building**.** Presence of diabetes and body mass index were added from the American College of Surgeons NSQIP. Univariate models were first constructed to assess relationships between the risk factor and the CDI incidence rate, then applied to a multivariate model. Selection criteria were eligibility for inclusion at a p value of 0.25 and retention at a p value of 0.05. Model validation was tested by a bootstrap sampling method and the results are provided. C-statistic for COLO complex 30-day model is 0.575 and for the HYST measure was 0.599. Hosmer-Lemeshow p-value for the COLO complex model was 0.64 and for the HYST model was 0.012.

Social risk factors were not specifically included due to data entry burden and a cited lack of evidence that supports the hypothesis that data collection of such would justify inclusion.

PANEL MEMBER 2: The c-statistic is low for both models and Hosmer-Lemeshow is significant for the hysterectomy model.

PANEL MEMBER 3: Variables in regression model initially identified through expert panels based on available data taking burden into account. Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but values not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on "optimism in the fit." No assessment made of stability of SIRs across bootstraps or change in ranking across significantly higher, significantly lower, although only 5% in higher and 6% in lower category for COLO where enough cases to categorize, and for HYST, 4% higher, 4% lower.

c-stats from risk adjustment models low (<.6), but that suggests risk adjustment explaining small amount of variance. However, given this, failure to consider how ranking changes given sizable variation in patient level predicted probabilities in 95% CI of risk model (3 times for COLO, 5 times for HYST) makes assessment of reliability limited.

Hosmer-Lemeshow p value 0.012 for HYST, suggesting problems of goodness of fit.

There is inadequate information presented on the stability of the measure to reasonable expected variation in counts of SSIs or the precision and stability of the risk adjustment model.

PANEL MEMBER 4: The c-statistics are pretty low suggesting the model is doing too much in terms of severity adjustment.

PANEL MEMBER 5: Risk model calibration (2b3.7) for SSI hyst shows potential issue with Hosmer Lemeshow test (p = 0.012). Developer does not discuss that, but cold be a sign of calibration issues.

VALIDITY: TESTING

17. **Validity testing level:** ☐ **Measure score** ☒ **Data element** ☐ **Both**
18. **Method of establishing validity of the measure score:**
    - ☐ **Face validity**
    - ☐ **Empirical validity testing of the measure score**
    - ☒ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

PANEL MEMBER 1: Seven states were presented showing sensitivity, specificity, positive predictive value, and negative predictive value for the COLO SSI with results of 74.9, 99.1, 95.8, and 93.5, respectively. For the HYST, the results are 80.7, 98.9, 92.6, and 96.9, respectively. Sensitivity is said to be lower due to a disincentive to report for facilities, further reflected by the negative predictive value.

PANEL MEMBER 2: Sensitivity/specificity analyses appear appropriate.

PANEL MEMBER 3: Chart review of numerator.

PANEL MEMBER 4: State based audits of clinical records in 8 states.

PANEL MEMBER 5: State validity testing

20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    PANEL MEMBER 1: Fair to acceptable results by this methodology

    PANEL MEMBER 2: Lower sensitivity is a concern.

    PANEL MEMBER 3: Estimates of sensitivity, specificity, positive and negative predicted value from chart review of sample.

    PANEL MEMBER 4: Very strong sensitivity and specificity. Sensitivity is lower than specificity (75% versus 99%), suggesting that this issue of identifying the correct events may be coming into play. The authors hypothesize this is related to CMS reporting requirements for COLO – seems like a testable hypothesis.

    PANEL MEMBER 5: Results are 90%+ for SSI hyst and SSI colo

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**

    NOTE that data element validation from the literature is acceptable.

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

    ☒ **Yes**

    ☐ **No**

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

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

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

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

☐ **Low** (NOTE: Should rate LOW if you believe that there 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.)

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

PANEL MEMBER 1: The seven/three state results analyzed for sensitivity, specificity, PPV, and NPV are presented. No other methodology is given.

PANEL MEMBER 2: Lower sensitivity rates and variation by state is a concern.

PANEL MEMBER 3: Sensitivity, specificity, PPV, NPV acceptable.

PANEL MEMBER 4: This is a very focused measure clearly based on a negative surgical event that hospitals and surgeons want to avoid. There seem to be some challenges converting this clinical knowledge into a standardized measure, such as exclusions and small Ns. As a result, I rated the measure moderate, not high.

PANEL MEMBER 5: No assessment of the validity or reliability of the risk adjusters was performed. I believe these are critical data elements, but would like to discuss with group.

ADDITIONAL RECOMMENDATIONS

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

PANEL MEMBER 1: The submission presented a sparse analysis for reliability and validity, mostly relying on the fact that it was based on a widely-used and valuable measure. It would have been nice to have seen a much more detailed review of the methodology. The risk adjustment strategy was good, but the dismissal of social risk factors, though conscious, was noted.

PANEL MEMBER 2: In the missing data section the developer states that 'rules present missing data'. However in the specifications section they indicate that the denominator excludes poor data quality and data errors. Would that not affect what would otherwise by counted as missing data?

PANEL MEMBER 3: The standing committee should discuss the basis for expert panel assessment of variables to be included in risk adjustment and extent to which variations in results across potential risk adjustment models should be presented.

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

2a1. Reliability – Specifications

Comments:

**Difficult to comment on because of the complexity of the methods described.

**There should be an age distribution result shown.

**Detailed specifications provided.

**Scientific methods panel raised issues mostly around colons.

**Consensus not reached. Personally do not have substantial concerns.

**None.

**Pondering the comments of the Scientific Methods group.

**I think the developer came back with solid testing no concerns.

**Reliability was estimated as the between- facility variance from a generalized linear mixed model divided by the total variance estimated from the same model. For COLO SSI, the mean reliability for facilities meeting the Minimum Precision Criteria (MPC) of at least one predicted event was 50.1% and 1,323 of 2,009 facilities had reliability exceeding 40%. For HYST SSI, the mean reliability for facilities meeting the MPC was 52.9% and 652 of 787 facilities meeting the MPC had reliability exceeding 40%.

2a2. Reliability – Testing

Comments:

**I defer to my research colleagues.

**No.

**Insufficient, reliability is 50% for COLO SSI and 53% for HYST SSI.

**Yes on colons would like to hear CDC and Methods panel discuss beforecommittee.

**Consensus not reached. Personally do not have substantial concerns.

**None.

**As above.

**I need more discussion with group.

**The measure has sufficient overall reliability.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences

Comments:

**I defer to my research colleagues.

**No.

**Moderate.

**No.

**Consensus not reached. Personally do not have substantial concerns.

**Moderate.

**As above.

**Sufficent.

**The SSI data used in this measure have been endorsed by NQF in a previous measure set and as described in 2b.2, the SMR, upon which the SIR is based, is a widely accepted method for summarizing mortality

experience. Therefore, we conclude the SIR measure has inherent face validity. However, we are undertaking validity studies beginning in July 2010.

**Concerns raised in this area by NQF staff deserve discussion, along with the discussion of reliability.

**The age range is still wide - can age differences be further refined.

**No concerns, small percent of excluded patients.

**No.

**Minimal but will likely require discussion given panel member comments.

**Moderate - missing data could present a problems.

**Agree w/ SM group - perhaps we should be concerned about "missing data".

**I need to hear the group on this there seems to be some vocal discussion about missing data and methodlogy.

**The Scientific Methods Panel did not reach consensus with the testing for the measure.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**The exclusions seem appropriate, but I would like to hear the discussion of the PSC.

**Was age adjusted for?

**Consensus panel some concerns with risk adjustment model.

**Risk adjustment done for both colons and hysterectomy.

**None.

**Exclusions seems appropriate.

**The Scientific Methods Panel did not reach consensus with the testing for the measure.

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.

Data Specifications and Elements

- The measure is constructed using data elements used in the generation of SSI data are routinely generated during care delivery. The NHSN analysis tool will automatically calculate SIRs.
- Some data elements are in defined fields in electronic sources. The developer noted that many of the data fields are available in electronic forms, and facilities may collect the denominator data and report it electronically via HL7 Clinical-Document Architecture file format. Some of the data fields for the numerator are available electronically, from sources such as laboratory reports, while others are found in plain language text.
    - However some of the data may not be available electronically.
    - The developer states there is no issues of missing data and data collection. SSI rates and SIR using the methodologies described above have been in use by hospitals participating in CDC surveillance systems since 1986

- This measure is not an eMeasure. At this time, the developer did not state any plan to specify the measure as an eMeasure.

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?
- Is the data collection strategy ready to be put into operational use?

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**Missing data appears to be a problem or shortcoming.

**Okay.

**Moderate, data elements routinely generated during care delivery, built into EHR, some data available electronically but not all.

**Can be either collected elctronically or on paper and submitted to NHSN.

**Agree with moderate feasibility prelim rating.

**Moderate - complex due to exclusions.

**No concerns.

**None.

**Data elements used in the generation of SSI data are routinely generated during care delivery. The NHSN analysis tool will automatically calculate SIRs.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

The measure is in four accountability program by Centers for Medicare and Medicaid Services listed below at the facility level and acute inpatient hospital setting:
1. Hospital Inpatient Quality Reporting Program
Purpose: To improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: Nationwide, currently covers all acute care hospitals with ICUs (approximately 3300).*

2. Prospective Payment System Exempt Cancer Hospital Quality Reporting Program
Purpose: To establish a quality reporting program for PPS-Exempt Cancer Hospital to improve health, improve care and lower cost (triple aims)of Medicare beneficiaries.

3. Hospital Value Based Purchasing
Purpose: To reward acute-care hospitals with incentive payments for the quality care provided to Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: Nationwide, includes all 2808 acute care facilities performing these procedures.

4. Hospital-acquired Condition(HAC) Reduction Program
Purpose: To provide an incentive for hospitals to reduce HACs
Geographic area and number and percentage of accountable entities and patients: Nationwide, includes all 3216 acute care facilities performing these procedures.

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

- Feedback on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided to us by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection control and surgical associations, and other personnel. Feedback was received via email regarding the extent of risk adjustment and the limitations. Different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users.

Additional Feedback:

- CMS also uses this measure for public reporting and payment programs.

***Questions for the Committee**:*

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

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

RATIONALE:

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

- SIRs following colon surgeries have been reduced by 6% and SSI SIRs following abdominal hysterectomies by 12% between 2015 and 2016, the most recent years with complete data.

- Abdominal Hysterectomy Surgical Site Infections (Facilities and National):
  - Facilities
    - 2015 Facility SIRS range-0.00 – 2.710 (median: 0.762)
    - 2016 Facility SIRS range-0.00 – 2.513 (median: 0.722)
  - National
    - National SSI HYST SIR in 2015 is 0.989 = 2,432 observed / 2,459.654 predicted SSIs
    - National SSI HYST SIR in 2016 is 0.868 = 2,138 observed / 2,462.289 predicted SSIs
    - Percent Change: 2016 v. 2015 12% decrease
- Colon Surgery Surgical Site Infections (Facilities and National):
  - Facilities
    - 2015 Facility SIRS range-0.00 – 2.399 (median: 0.783).
    - 2016 Facility SIRS range-0.00 – 2.170 (median: 0.781).
  - National
    - National SSI COLO SIR in 2015 is 0.989 = 8010 observed / 8,102.668 predicted SSIs
    - National SSI COLO SIR in 2016 is 0.931 = 7,960 observed / 8,553.309 predicted SSIs
    - Percent Change: 2016 v. 2015 6% decrease

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 noted there is potential for underreporting SSI events and/or the SIRs may be miscalculated and have an SIR that is higher than actual. However the developer notes the NHSN reporting tool includes business logic to minimize misclassification of SSI.

Potential harms

- As noted above, there is potential for underreporting SSI event and/or the SIRs may be miscalculated and have an SIR that is higher than actual

Additional Feedback:

- N/A

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?

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

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

RATIONALE:

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

4a1. Use - Accountability and Transparency
Comments:

** Already being used in multiple federal accountability programs.

**I think the cases reported seems too limited for the population and incidence by age.

**High, publically reported by many organizations including CMS, NHSN runs analysis, hospitals use SIR analysis for prevention activities.

**Publically reported not clear how much progress is being made in colons.

**Agree with prelim rating.

**High.

**Used in payment programs.

**Sufficient.

**NHSN users can run monthly analysis reports within NHSN to view their SIR data. On an annual basis, NHSN publishes national and state- level SIRs in the National and State HAI Progress Report. State health departments that perform validation of SSI data reported to NHSN, provide feedback to facilities.

4b1. Usability – Improvement
Comments:

** Benefits seem to outweigh any unintended consequences.

**Limited.

** High, results improving, but risk of underreporting.

** Not sure yet.

**Minimal apparent unintended consequences.

**None.

**Quite uncertain as to whether poor performing sites are "using" and making any improvements.

**Pass.

** The measure is in four accountability programs by Centers for Medicare and Medicaid Services at the facility level and acute inpatient hospital setting.

Criterion 5: Related and Competing Measures

Related or competing measures
Related measures:

3025 : Ambulatory Breast Procedure Surgical Site Infection (SSI) Outcome Measure

Harmonization
The developer notes that the settings differ for 3025 and 0753. 3025 is performed at ambulatory surgery centers whereas 0753 is performed at inpatient facility level. In addition, these two measure target

populations have potential difference in SSI risk as their comorbidities, types of procedures performed, and length of time cared for in a healthcare facility are inherently different. Risk modeling has been performed for both measures, with different models developed based on procedure and facility type.

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

5. Related and Competing

Comments

**None.

**No.

**Related measure 3025: Ambulatory Breast Procedure Surgical Site Infection (SSI) Outcome Measure, but different populations in different settings.

**Not for inpatient.

** 3025: Ambulatory measure.

**No important competing.

**Not specifically conflicting.

**Related measure but not competing -- ASC vs Hospital.

** The developer notes that the settings differ for 3025 and 0753. 3025 is performed at ambulatory surgery centers whereas 0753 is performed at inpatient facility level. In addition, these two measure target populations have potential difference in SSI risk as their comorbidities, types of procedures performed, and length of time cared for in a healthcare facility are inherently different. Risk modeling has been performed for both measures, with different models developed based on procedure and facility type.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/19

- No NQF Members have submitted support/non-support choices as of this date.

Brief Measure Information

NQF #: 0753

Corresponding Measures:

De.2. Measure Title: American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure

Co.1.1. Measure Steward: Centers for Disease Control and Prevention

De.3. Brief Description of Measure: Facility adjusted Standardized Infection Ratio (SIR) and Adjusted Ranking Metric (ARM) for deep incisional and organ/space Surgical Site Infections (SSI) at the primary incision site among adult patients aged >= 18 years as reported through the CDC National Health and Safety Network (NHSN).

1b.1. Developer Rationale: It is envisioned the use of this measure will promote SSI prevention activities which will lead to improved patient outcomes including reduction of avoidable medical costs, and patient morbidity and mortality.

S.4. Numerator Statement: Deep incisional primary (DIP) and organ/space SSIs during the 30-day postoperative period among patients = 18 years of age, who undergo inpatient colon surgeries or abdominal hysterectomies. SSIs will be identified before discharge from the hospital, upon readmission to the same hospital, or during outpatient care or admission to another hospital (post-discharge surveillance).

Numerator Exclusion SSI events with PATOS* field = yes.

Infection present at time of surgery (PATOS): PATOS denotes that there is evidence of an infection or abscess at the start of or during the index surgical procedure (in other words, it is present preoperatively). PATOS is a YES/NO field on the SSI Event form. PATOS does not apply if there is a period of wellness between the time of a preoperative condition and surgery. The evidence of infection or abscess must be noted/documented intraoperatively in an operative note or report of surgery. Only select PATOS = YES if it applies to the depth of SSI that is being attributed to the procedures (e.g., if a patient has evidence of an intraabdominal infection at the time of surgery and then later returns with an organ/space SSI the PATOS field would be selected as a YES. If the patient returned with a superficial or deep incisional SSI the PATOS field would be selected as a NO). The patient does not have to meet the NHSN definition of an SSI at the time of the primary procedure but there must be notation that there is evidence of an infection or abscess present at the time of surgery. PATOS is not necessarily diagnosis driven.

S.6. Denominator Statement: An NHSN Operative Procedure is a procedure:

• that is included in the ICD-10-PCS or CPT NHSN operative procedure code mapping. And

• takes place during an operation where at least one incision (including laparoscopic approach and cranial Burr holes) is made through the skin or mucous membrane, or reoperation via an incision that was left open during a prior operative procedure And

• takes place in an operating room (OR), defined as a patient care area that met the Facilities Guidelines Institute's (FGI) or American Institute of Architects' (AIA) criteria for an operating room when it was constructed or renovated. This may include an operating room, C-section room, interventional radiology room, or a cardiac catheterization lab.

Exclusions: Otherwise eligible procedures that are assigned an ASA score of 6 are not eligible for NHSN SSI surveillance.

Using multivariable logistic regression models for colon surgeries and abdominal hysterectomies, the predicted number of SSIs is obtained. These predicted numbers are summed by facility and surgical procedure and used as the denominator of this measure (see also 2a.8).

S.8. Denominator Exclusions: Denominator data are excluded from the SSI measure due to various reasons related to data quality, data outlier and data errors. The complete list of universal exclusion criteria applied to denominator are listed in the SSI section of the SIR guide that is referenced above. These exclusions include but are not limited to procedures associated with SSI events where the PATOS = yes, and those with ASA Class VI (6). The measure specific denominator exclusions for the Complex 30-day SSI, are off plan colon and abdominal hysterectomy procedures, procedures performed on persons under the age of 18, and procedure performed on an outpatient basis. .

Note: Under the 2015 baseline, both primarily closed procedures and those that are not closed primarily are included in the denominator data.Persons under the age of 18, those having a procedure performed on an outpatient basis, procedures associated with SSI events where the PATOS = yes, those with ASA Class VI (6) are excluded.

Note: Both primarily closed procedures and those that are not closed primarily are included in the denominator data.

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

S.20. Level of Analysis: Facility, Other, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2012 **Most Recent Endorsement Date:** Jan 17, 2012

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?

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

0753_Evidence_MSF5.0_Data-636687169504749334.doc,NQF_evidence_-Importance-_Final.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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*)**:** 0753

Measure Title: American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 8/1/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

☒ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Electronic health records, laboratory, other, paper medical records

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Previously identified interventions have been shown to reduce the incidence of surgical site infections including, but not limited to use of sterile technique, avoidance of preoperative shaving of the operative site, preoperative decontamination of the surgical site, administration of preoperative prophylactic antibiotics within a prescribed timeframe, maintaining glycemic control in diabetic patients, and providing an increased inspired fraction of oxygen to the patient during and immediately following surgery.

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

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

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

Berríos-Torres, SI. et al., Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection. JAMA Surg, 152(8): (2017):784-791.

Berrios-Torres, SI, et al, screened 5759 titles and abstracts, of which 896 underwent full-text review by 2 independent reviewers. After exclusions, 170 studies were extracted into evidence tables, appraised, and synthesized. Based on these studies, recommendations for preventing SSI were:

- Before surgery, patients should shower or bathe (full body) with soap (antimicrobial or nonantimicrobial) or an antiseptic agent on at least the night before the operative day.
- Antimicrobial prophylaxis should be administered only when indicated based on published clinical practice guidelines and timed such that a bactericidal concentration of the agents is established in the serum and tissues when the incision is made.

- Skin preparation in the operating room should be performed using an alcohol-based agent unless contraindicated.
- For clean and clean-contaminated procedures, additional prophylactic antimicrobial agent doses should not be administered after the surgical incision is closed in the operating room, even in the presence of a drain.
- Topical antimicrobial agents should not be applied to the surgical incision.
- During surgery, glycemic control should be implemented using blood glucose target levels less than 200 mg/dL, and normothermia should be maintained in all patients.
- Increased fraction of inspired oxygen should be administered during surgery and after extubation in the immediate postoperative period for patients with normal pulmonary function undergoing general anesthesia with endotracheal intubation.
- Transfusion of blood products should not be withheld from surgical patients as a means to prevent SSI.

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

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

☒ Clinical Practice Guideline recommendation (with evidence review)

An update of The Center for Disease Control and Prevention's Guideline for the Prevention of Surgical Site Infection has been published since the last submission.

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	

Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

It is envisioned the use of this measure will promote SSI prevention activities which will lead to improved patient outcomes including reduction of avoidable medical costs, and patient morbidity and mortality.

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.*

When SIRs are compared over time, assessment of performance can be made. CDC has demonstrated significant performance gaps in SIRs across facilities. See below:

Surgical Site Infection (SSI) – Abdominal Hysterectomy

2015-2016 SSI- HYST: SSIs included are those classified as deep incisional or organ/space infections following 30 days after inpatient procedures in adults 18 years and older that occurred in 2015 and 2016 with primary and other than primary skin closure technique, detected during admission as the surgical procedure or upon readmission to the same facility or different facility or through post discharge surveillance. (Complex 30-day SSI model)

2015:

o 3,426 facilities reporting, 303,361 in-plan, inpatient HYST procedures in adults 18 years and older (with no considerations to the SIR exclusion criteria)

o Facility SIRs range from 0.00 – 2.710 (median: 0.762) The facility SIR distribution is 5% at the minimum and 95% at the maximum

2016:

o 3,447 facilities reporting, 300,483 in-plan, inpatient HYST procedures in adults 18 years and older (with no considerations to the SIR exclusion criteria)

o Facility SIRs range from 0.00 – 2.513 (median: 0.722) The facility SIR distribution is 5% at the minimum and 95% at the maximum

National SSI HYST SIR in 2015 is 0.989 = 2,432 observed / 2,459.654 predicted SSIs

National % change vs. baseline in 2015 shows 1.1% difference

National SSI HYST SIR in 2016 is 0.868 = 2,138 observed / 2,462.289 predicted SSIs

National % change vs. baseline in 2016 is 13%

Percent Change: 2016 v. 2015 12% decrease

Surgical Site Infection (SSI) – Colon Surgery

2015-2016 SSI- HYST: SSIs included are those classified as deep incisional or organ/space infections following 30 days after inpatient procedures in adults 18 years and older that occurred in 2015 and 2016 with primary and other than primary skin closure technique, detected during admission as the surgical procedure or upon readmission to the same facility or different facility or through post discharge surveillance.. (Complex 30-day SSI model)

2015:

o 3,433 facilities reporting, 306,536 in-plan, inpatient COLO procedures in adults 18 years and older (with no considerations to the SIR exclusion criteria)

o Facility SIRs range from 0.00 – 2.399 (median: 0.783). The facility SIR distribution is 5% at the minimum and 95% at the maximum

2016:

o 3,461 facilities reporting, 321, 535 in-plan COLO procedures in adults 18 years and older (with no considerations to the SIR exclusion criteria)

o Facility SIRs range from 0.00 – 2.170 (median: 0.781). The facility SIR distribution is 5% at the minimum and 95% at the maximum

National SSI COLO SIR in 2015 is 0.989 = 8010 observed / 8,102.668 predicted SSIs

National % change vs. baseline in 2015 is 1.1%

National SSI COLO SIR in 2016 is 0.931 = 7,960 observed / 8,553.309 predicted SSIs

National % change vs. baseline in 2016 is 6.9%

Percent Change: 2016 v. 2015 6% decrease

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.

The data presented in the reports display the status of HAI in the United States over time and currently.

The Healthcare-associated Infections in the United States, 2006-2016: A Story of Progress located here: https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html

The 2015 National and State Healthcare-associated Infection Data Report: https://www.cdc.gov/hai/surveillance/data-reports/2015-HAI-data-report.html

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.*

No studies provide evidence of a direct relationship between social risk and HAIs. Instead, they provide evidence that social risk factors are associated with an increased risk of chronic disease conditions, suboptimal care for those conditions, compromised functional status, exposure to nursing homes, and colonization with bacterial pathogens. While these associations may be meaningful, they do not establish a direct relationship between social risk factors and HAIs. Some evidence to the contrary is available. For example, empirical data analyzed by CDC, specifically surgical site infection (SSI) data that hospitals report to NHSN, have yielded findings that SSI risk is lower among older patients compared to younger patients for some surgical procedures: http://www.cdc.gov/nhsn/pdfs/pscmanual/ssi_modelpaper.pdf Until more compelling evidence of a direct relationship between social risk and HAIs is available, it would be premature to adjust for social risk factors in the clinical quality measures that CDC reports.

Certain patient-related factors have been associated with an increased risk of SSI, using the Complex 30-day SSI model, including for COLO: advanced age, [1] [2], , American Society of Anesthesiologists' physical status classification (ASA) >2, [2, 3] diabetes, gender, obesity, procedure closure technique and procedures performed at oncology facilities vs. those performed at non oncology facilities.

For HYST: younger age, increasing ASA classification, obesity, diabetes, and procedures performed at oncology facilities vs. those performed at non-oncology facilities. .

Oncology facilities are those facilities designated in NHSN as providing services exclusively to oncology patients

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

1. Berríos-Torres, SI. et al., Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection. JAMA Surg, 152(8): (2017):784-791.

2. Neumayer, L., et al., Multivariable predictors of postoperative surgical site infection after general

and vascular surgery: results from the patient safety in surgery study. J Am Coll Surg, 2007. 204(6): p.

1178-87.

3. Yi Mu, Jonathan R. Edwards, Teresa C. Horan, Sandra I. Berrios-Torres, Scott K. Fridkin , Improving Risk-Adjusted Measures of Surgical Site Infection for the National Healthcare Safety Network: Infection Control and Hospital Epidemiology, Vol. 32, No. 10 (October 2011), pp. 970¬-986

4. NHSN HAI Progress Report: https://www.cdc.gov/hai/pdfs/progress-report/hai-progress-report.pdf

5. Deverick J. Anderson, MD, MPH; Kelly Podgorny, DNP, MS, RN; Sandra I. Berríos-Torres, MD; Dale W. Bratzler, DO, MPH; E. Patchen Dellinger, MD; Linda Greene, RN, MPS,CIC; Ann-Christine Nyquist, MD, MSPH; Lisa Saiman, MD, MPH; Deborah S. Yokoe, MD, MPH; Lisa L. Maragakis, MD, MPH; Keith S. Kaye, MD, MPH, Strategies to Prevent Surgical Site Infections in Acute Care Hospitals: 2014 Update. Infection Control and Hospital Epidemiology, Vol. 35, No. 6 (June 2014), pp. 605-627

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

Elderly

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

https://www.cdc.gov/nhsn/acute-care-hospital/ssi/index.html

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** icd10-pcs-pcm-nhsn-opc.xlsx

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

No, this is not an instrument-based measure **Attachment:**

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.

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.

There has been an update to the 2015 baseline model for the Complex 30-day SSI model (CMS model). Link to the SIR Guide in which the new model details are published in S.2b The national aggregate data or baseline data previously used in the calculation of the risk adjusted measure was updated due to the growth in

surveillance of the HAI data over time. Some of the important changes include the difference in patient mix over time, the increase in the incidence of infection, the increase in the number of reporting facilities and the changes in some surveillance definitions.

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).

Deep incisional primary (DIP) and organ/space SSIs during the 30-day postoperative period among patients = 18 years of age, who undergo inpatient colon surgeries or abdominal hysterectomies. SSIs will be identified before discharge from the hospital, upon readmission to the same hospital, or during outpatient care or admission to another hospital (post-discharge surveillance).

Numerator Exclusion SSI events with PATOS* field = yes.

Infection present at time of surgery (PATOS): PATOS denotes that there is evidence of an infection or abscess at the start of or during the index surgical procedure (in other words, it is present preoperatively). PATOS is a YES/NO field on the SSI Event form. PATOS does not apply if there is a period of wellness between the time of a preoperative condition and surgery. The evidence of infection or abscess must be noted/documented intraoperatively in an operative note or report of surgery. Only select PATOS = YES if it applies to the depth of SSI that is being attributed to the procedures (e.g., if a patient has evidence of an intraabdominal infection at the time of surgery and then later returns with an organ/space SSI the PATOS field would be selected as a YES. If the patient returned with a superficial or deep incisional SSI the PATOS field would be selected as a NO). The patient does not have to meet the NHSN definition of an SSI at the time of the primary procedure but there must be notation that there is evidence of an infection or abscess present at the time of surgery. PATOS is not necessarily diagnosis driven.

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).

Colon surgeries: Defined by the ICD-9-CM procedure codes that comprise the NHSN colon surgery category for that program, or the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

Abdominal hysterectomy: Defined by the ICD-9-CM procedure codes that comprise the NHSN abdominal hysterectomy category for that program, or the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

Inpatient: A patient for whom the discharge date is at least one day later than the admission date

Adult: A person =18 years of age

A deep incisional SSI must meet one of the following criteria:

The date of event for infection occurs within 30 days after the NHSN operative procedure (where day 1 = the procedure date)

AND

involves deep soft tissues of the incision (e.g., fascial and muscle layers)

AND

patient has at least one of the following:

a. purulent drainage from the deep incision.

b. a deep incision that spontaneously dehisces, or is deliberately opened or aspirated by a surgeon, attending physician** or other designee and organism is identified by a culture or non-culture based microbiologic testing method which is performed for purposes of clinical diagnosis or treatment (e.g., not Active Surveillance Culture/Testing (ASC/AST) or culture or non-culture based microbiologic testing method is not performed

AND

patient has at least one of the following signs or symptoms: fever(>38°C); localized pain or tenderness. A culture or non-culture based test that has a negative finding does not meet this criterion.

c. an abscess or other evidence of infection involving the deep incision that is detected on gross anatomical or histopathologic exam, or imaging test

** The term attending physician for the purposes of application of the NHSN SSI criteria may be interpreted to mean the surgeon(s), infectious disease, other physician on the case, emergency

An organ/space SSI involves any part of the body deeper than the fascial/muscle layers, that is opened or manipulated during the operative procedure. The table below lists the specific sites that must be used to differentiate organ/space SSI. Specific sites are assigned to organ/space SSI to further identify the location of the infection. Specific sites of organ/space have specific criteria which must be met in order to qualify as an NHSN event. These criteria are in addition to the general criteria for NHSN organ/space SSI.

Specific sites of Organ/space events available for COLO and HYST.

COLO - Colon surgery GIT - Gastrointestinal tract

IAB - Intraabdominal, not specified elsewhere

OREP - Other infection of the male or female reproductive tract

USI - Urinary System Infection

HYST - Abdominal hysterectomy IAB - Intraabdominal, not specified elsewhere

OREP - Other infection of the male or female reproductive tract

VCUF - Vaginal cuff infection

An organ/space SSI must meet one of the following criteria:

Date of event for infection occurs within 30 days after the NHSN operative procedure (where day 1 = the procedure date)

AND

infection involves any part of the body deeper than the fascial/muscle layers, that is opened or manipulated during the operative procedure

AND

patient has at least one of the following:

a. purulent drainage from a drain that is placed into the organ/space (e.g., closed suction drainage system, open drain, T-tube drain, CT guided drainage)

b. organisms are identified from an aseptically-obtained fluid or tissue in the organ/space by a culture or non-culture based microbiologic testing method which is performed for purposes of clinical diagnosis or treatment (e.g., not Active Surveillance Culture/Testing (ASC/AST).

c. an abscess or other evidence of infection involving the organ/space that is detected on gross anatomical or histopathologic exam, or imaging test evidence suggestive of infection.

AND

meets at least one criterion for a specific organ/space infection site listed in COLO and HYST tables above.

These criteria are found in the Surveillance Definitions for Specific Types of Infections chapter 17.

REPORTING INSTRUCTIONS:

Multiple tissue levels are involved in the infection: The type of SSI (superficial incisional, deep incisional, or organ/space) reported should reflect the deepest tissue layer involved in the infection during the surveillance period. The date of event should be the date that the patient met criteria for the deepest level of infection:

a. Report infection that involves the organ/space as an organ/space SSI, whether or not it also involves the superficial or deep incision sites.

b. Report infection that involves the superficial and deep incisional sites as a deep incisional SSI.

c. If an SSI started as a deep incisional SSI on day 10 of the SSI surveillance period and then a week later, (day 17 of the SSI surveillance period) meets criteria for an organ space SSI the date of event would be the date of the organ space SSI.

Patient Specific Data:

Procedure/SSI Complex 30-Day Model- 2015 Baseline

COLO HYST

Diabetes Diabetes

ASA Score ASA Score

Age Age

Gender BMI

BMI Cancer Hospital

Closure technique

Cancer HospitalColon surgeries: Defined by the ICD-10-PCS procedure codes that comprise the NHSN colon surgery category for that program, or the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

Abdominal hysterectomy: Defined by the ICD-10-PCS procedure codes that comprise the NHSN abdominal hysterectomy category for that program, or the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

Inpatient: A patient for whom the discharge date is at least one day later than the admission date

Adult: A person =18 years of age

A deep incisional SSI must meet one of the following criteria:

The date of event for infection occurs within 30 days after the NHSN operative procedure (where day 1 = the procedure date)

AND

involves deep soft tissues of the incision (e.g., fascial and muscle layers)

AND

patient has at least one of the following:

a. purulent drainage from the deep incision.

b. a deep incision that spontaneously dehisces, or is deliberately opened or aspirated by a surgeon, attending physician** or other designee

AND

organism is identified by a culture or non-culture based microbiologic testing method which is performed for purposes of clinical diagnosis or treatment (e.g., not Active Surveillance Culture/Testing (ASC/AST) or culture or non-culture based microbiologic testing method is not performed

AND

patient has at least one of the following signs or symptoms: fever(>38°C); localized pain or tenderness. A culture or non-culture based test that has a negative finding does not meet this criterion.

c. an abscess or other evidence of infection involving the deep incision that is detected on gross anatomical or histopathologic exam, or imaging test

** The term attending physician for the purposes of application of the NHSN SSI criteria may be interpreted to mean the surgeon(s), infectious disease, other physician on the case, emergency

An organ/space SSI involves any part of the body deeper than the fascial/muscle layers that is opened or manipulated during the operative procedure. The table below lists the specific sites that must be used to differentiate organ/space SSI. Specific sites are assigned to organ/space SSI to further identify the location of the infection. Specific sites of organ/space have specific criteria which must be met in order to qualify as an NHSN event. These criteria are in addition to the general criteria for NHSN organ/space SSI.

Specific sites of Organ/space events available for COLO and HYST.

COLO - Colon surgery

GIT - Gastrointestinal tract

IAB - Intraabdominal, not specified elsewhere

OREP - Other infection of the male or female reproductive tract

USI - Urinary System Infection

HYST - Abdominal hysterectomy

IAB - Intraabdominal, not specified elsewhere

OREP - Other infection of the male or female reproductive tract

VCUF - Vaginal cuff infection

An organ/space SSI must meet one of the following criteria:

Date of event for infection occurs within 30 days after the NHSN operative procedure (where day 1 = the procedure date)

AND

infection involves any part of the body deeper than the fascial/muscle layers, that is opened or manipulated during the operative procedure

AND

patient has at least one of the following:

a. purulent drainage from a drain that is placed into the organ/space (e.g., closed suction drainage system, open drain, T-tube drain, CT guided drainage)

b. organisms are identified from an aseptically-obtained fluid or tissue in the organ/space by a culture or non-culture based microbiologic testing method which is performed for purposes of clinical diagnosis or treatment (e.g., not Active Surveillance Culture/Testing (ASC/AST).

c. an abscess or other evidence of infection involving the organ/space that is detected on gross anatomical or histopathologic exam, or imaging test evidence suggestive of infection.

AND

meets at least one criterion for a specific organ/space infection site listed in COLO and HYST tables above.

These criteria are found in the Surveillance Definitions for Specific Types of Infections chapter 17.

REPORTING INSTRUCTIONS:

Multiple tissue levels are involved in the infection: The type of SSI (superficial incisional, deep incisional, or organ/space) reported should reflect the deepest tissue layer involved in the infection during the surveillance period. The date of event should be the date that the patient met criteria for the deepest level of infection:

a. Report infection that involves the organ/space as an organ/space SSI, whether or not it also involves the superficial or deep incision sites.

b. Report infection that involves the superficial and deep incisional sites as a deep incisional SSI.

c. If an SSI started as a deep incisional SSI on day 10 of the SSI surveillance period and then a week later, (day 17 of the SSI surveillance period) meets criteria for an organ space SSI the date of event would be the date of the organ space SSI.

Patient Specific Data:

Procedure/SSI Complex 30-Day Model- 2015 Baseline

Complex 30-day SSI Model: COLO

Diabetes

ASA Score

Age

Gender

BMI

Cancer hospital

Closure technique

Complex 30-day SSI Model: HYST

Diabetes

ASA Score

Age

BMI

Cancer hospital

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

An NHSN Operative Procedure is a procedure:

• that is included in the ICD-10-PCS or CPT NHSN operative procedure code mapping. And

• takes place during an operation where at least one incision (including laparoscopic approach and cranial Burr holes) is made through the skin or mucous membrane, or reoperation via an incision that was left open during a prior operative procedure And

• takes place in an operating room (OR), defined as a patient care area that met the Facilities Guidelines Institute's (FGI) or American Institute of Architects' (AIA) criteria for an operating room when it was constructed or renovated. This may include an operating room, C-section room, interventional radiology room, or a cardiac catheterization lab.

Exclusions: Otherwise eligible procedures that are assigned an ASA score of 6 are not eligible for NHSN SSI surveillance.

Using multivariable logistic regression models for colon surgeries and abdominal hysterectomies, the predicted number of SSIs is obtained. These predicted numbers are summed by facility and surgical procedure and used as the denominator of this measure (see also 2a.8).

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).

Data required to calculate the denominator:

1) Data for each operative procedure

Colon surgeries: Defined by the ICD-10-PCS procedure codes that comprise the NHSN colon surgery category for that program, and or the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

Abdominal hysterectomy: Defined by the ICD-10-PCS procedure codes that comprise the NHSN abdominal hysterectomy category for that program, or and the corresponding set of CPT procedure codes used in ACS/NSQIP for that program (see Appendix 1).

2) Parameter estimates for operative procedure-specific logistic regression models are needed to calculate the predicted number of SSIs. See pages 29 of the SIR guide, 2a.15 attachment.

3) Patient Specific Data: Procedure/SSI Complex 30-Day Model- 2015 Baseline

Complex 30-day SSI Model: COLO

Diabetes

ASA Score

Age

Gender

BMI

Cancer hospital

Closure technique

Complex 30-day SSI Model: HYST

Diabetes

ASA Score

Age

BMI

Cancer hospital

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

Denominator data are excluded from the SSI measure due to various reasons related to data quality, data outlier and data errors. The complete list of universal exclusion criteria applied to denominator are listed in the SSI section of the SIR guide that is referenced above. These exclusions include but are not limited to procedures associated with SSI events where the PATOS = yes, and those with ASA Class VI (6). The measure specific denominator exclusions for the Complex 30-day SSI, are off plan colon and abdominal hysterectomy procedures, procedures performed on persons under the age of 18, and procedure performed on an outpatient basis. .

Note: Under the 2015 baseline, both primarily closed procedures and those that are not closed primarily are included in the denominator data.Persons under the age of 18, those having a procedure performed on an

outpatient basis, procedures associated with SSI events where the PATOS = yes, those with ASA Class VI (6) are excluded.

Note: Both primarily closed procedures and those that are not closed primarily are included in the denominator data.

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.)*

Age (person is under 18)

Date of admission and date discharge on the same calendar day

Procedures associated with a PATOS = yes SSI event

ASA Class (6)

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.)*

None

If desired by an implementing organization or agency, race and ethnicity information could be added to data collection to allow for post-hoc stratification to identify disparities by these groupings. Risk adjustment based on these variables is not proposed.None

If desired by an implementing organization or agency, race and ethnicity information could be added to data collection to allow for post-hoc stratification to identify disparities by these groupings. Risk adjustment based on these variables is not proposed.

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

Other

If other: The measure reports the individual adjusted Standardized Infection Ratio (SIR) for colon surgeries and abdominal hysterectomies for each facility during the specified reporting period. SIR is an indirect standardization method for summarizing healthcare associated infection (HAI) experience across any number of stratified groups of data. Because the facility SIR has lower precision for facilities with few expected events relative to the number of procedures performed, i.e. low reliability, empirical Bayes techniques are used to derive the final reported SIR or reliability-adjusted SIR.

S.12. Type of score:

Other

If other: Adjusted Ratio: The reliability adjusted SIR is the reliability adjusted number of SSIs divided by the expected number of SSIs. The reliability adjustment for each facility is based on procedure volume.

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 = Lower 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.)*

An SIR <1.0 indicates that the number of SSIs was fewer than expected for that facility, whereas an SIR >1.0 indicates that the number of SSIs was more than expected, given the patients treated.

An ARM <1.0 indicates that the number of SSIs was fewer than expected for that facility, whereas an ARM >1.0 indicates that the number of SSIs was more than expected, given the patients treated.

The SIR is calculated as follows:

1. Identify the number of SSIs for each procedure

2. Total these numbers for an observed number of SSIs

3. Obtain the predicted number of SSIs for each procedure by multiplying the observed number of procedures by the corresponding SSI rates for each procedure from a standard population (as reflected in the regression models, see section 2b.3 Testing Results)

4. Sum the number of predicted SSIs for each procedure in the measurement time period.

5. Divide the total number of observed SSIs ("2" above) by the "predicted" number of SSIs ("4" above).

6. Result = SIR

An ARM <1.0 indicates that the number of SSIs was fewer than expected for that facility, whereas an ARM >1.0 indicates that the number of SSIs was more than expected, given the patients treated.

The SIR is calculated as follows:

1. Identify the number of SSIs for each procedure

2. Total these numbers for an observed number of SSIs

3. Obtain the predicted number of SSIs for each procedure by multiplying the observed number of procedures by the corresponding SSI rates for each procedure from a standard population (as reflected in the regression models, see section 2b.3 Testing Results)

4. Sum the number of predicted SSIs for each procedure in the measurement time period.

5. Divide the total number of observed SSIs ("2" above) by the "predicted" number of SSIs ("4" above).

6. Result = SIR

The reliability ARM is calculated as follows:

1. Obtain the adjusted number of observed SSI by using a Bayesian posterior distribution constructed through Monte Carlo Markov Chain sampling which results from a Bayesian random effects model.

2. Sum these adjusted number of observed SSI by hospital for the adjusted observed SSIs total.

3. For every patient undergoing the operative procedure in the period, calculate the probability of SSI using the patient data and parameter estimates of the factors in the applicable model.

4. Sum the probabilities by hospital to obtain the total expected number of SSIs.

5. Divide the total number of adjusted observed SSIs by the total number of expected SSIs for the resulting ARM.

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.

No sampling

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.

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

If other, please describe in S.18.

Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

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.

Data will be reported using the formats in the following forms:

1) NHSN SSI Event form (CDC 57.120)

2) NHSN Denominator for Procedure form (CDC 57.121)

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

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

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

Facility, Other, Population : Regional and State

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

Inpatient/Hospital

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.)*

2. Validity – See attached Measure Testing Submission Form

ICD-9-cmCODES.xlsx,SSI__NQF_testing_Final_submit_11-9-18.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.

No

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*)**:** 0753
Measure Title: American College of Surgeons – Centers for Disease Control and Prevention (ACS-CDC) Harmonized Procedure Specific Surgical Site Infection (SSI) Outcome Measure
Date of Submission: 7/31/2018
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant

difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☒ other: National Healthcare Safety Network

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*). National Healthcare Safety Network (NHSN)

1.3. What are the dates of the data used in testing? January 1-December 31, 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: Population: Regional and State	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*) The SIR for COLO SSI is based on 3,388 facilities. The SIR for HYST SSI is based on 3,250 facilities.

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*) Estimation of the SIR for COLO SSI included 304,173 patients with 8,266 events. Estimation of the SIR for HYST SSI included 304,735 patients with 2,515 events.

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. Facilities are attributed an SIR only if they meet the Minimum Precision Criteria (MPC) which requires a facility to have at least one predicted event from the risk-adjustment model. 2,009 facilities met the MPC for COLO SSI and 787 facilities met the MPC for HYST SSI.

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. To date, evidence has not substantiated the relationship between socioeconomic status and SSI. Please see response to 2b3.4b

---

2a2. RELIABILITY TESTING

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

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

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

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

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The SSI data used in this measure have been endorsed by NQF in a previous measure set and as described in 2b.2, the SMR, upon which the SIR is based, is a widely accepted method for summarizing mortality experience. Therefore, we conclude the SIR measure has inherent face validity. However, we are undertaking validity studies beginning in July 2010.

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

Reliability was estimated as the between- facility variance from a generalized linear mixed model divided by the total variance estimated from the same model. For COLO SSI, the mean reliability for facilities meeting the Minimum Precision Criteria (MPC) of at least one predicted event was 50.1% and 1,323 of 2,009 facilities had reliability exceeding 40%. For HYST SSI, the mean reliability for facilities meeting the MPC was 52.9% and 652 of 787 facilities meeting the MPC had reliability exceeding 40%.

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?*)

The reliability of a facility is related to the number of procedures conducted: facilities with a low number of procedures tend to have lower reliability. The MPC is intended to remove facilities that lack sufficient procedures required to reliably estimate an SIR. Overall, the mean reliability is adequate. Around one-third of facilities that met the MPC had reliability below the commonly-used 40% threshold for COLO SSI although

most facilities meeting the MPC had reliability above the 40% threshold for HYST SSI. Our conclusion is that the measure has sufficient overall reliability.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☐ **Performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

State validity testing and CMS validity testing performed tested Classification Error.

This reliability testing consisted of identifying patients that had undergone an abdominal hysterectomy or colon surgery and were therefore at risk for SSI. States' methodologies and sampling practices varied but in general, auditors reviewed postoperative medical records for signs and symptoms and determinations made as to whether the patient met criteria for NHSN SSI. Comparisons were then made to data reported to the NHSN and accuracy of SSI reporting was computed.

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

Between 2013 and 2017, seven states conducted validation studies of COLO SSI data within their states. No separate HYST SSI validations have been performed by states.

Mean measurements identified (ranges) were as follows:

Sensitivity: 74.9% (59.8-90.1)

Specificity: 99.1 % (98.7-100)

Positive Predictive Value: 95.8% (91.7-100)

Negative Predictive Value: 93.5% (85.3-97.2)

SSI COLO

	Year	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Colorado	2013	68	98.7	91.7	93.5
Maine	2014	87	99.1	97	95.8
Utah	2015	75.8	99.4	97.3	93.9
California	2014	69			

	Year	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
New York	2014	59.8	99.5	98.1	85.3
New Hampshire	2015	81.4	100	100	88.6
Wisconsin	2016	90.1	98.4	94.1	97.2
Overall		**74.9**	**99.1**	**95.8**	**93.5**

SSI HYST

	Year	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
New Hampshire	**2015**	100	88.9	94.4	100
New York	2014	75.4	99.1	91.5	96.9
Mississippi	2016	80	99.0	93.3	96.7
Overall		**80.7**	**98.9**	**92.6**	**96.9**

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?*)

State validation audits show that COLO SSIs reported by facilities meet the requirements of the NHSN SSI measure. Results for sensitivity are lower, however, there are disincentives for facilities to report COLO SSIs to NHSN. This is because COLO SSI SIRs are used by the Centers for Medicare and Medicaid Services (CMS) to determine financial reimbursements for facilities participating in CMS Quality Performance Programs. Therefore, the lower sensitivity may be a reflection of reticence to report rather than reliability of the measure. This issue is also reflected in the negative predictive value.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

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*) CDC adopted the American College of Surgeons' process of exclusion from SSI counts, those procedures identified with infections Present at the Time of Surgery (PATOS).

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*) 6669 (2.15%) of patients were excluded from COLO procedure and 4241 (1.37%) of patients were excluded from HYST procedures.

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*)

_____

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 2b5**.**

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model** with 7 risk factors for COLO and 5 risk factors for HYST.

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

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

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

Multiple logistic regression model was used to build risk adjusted models.

Risk factors for SSI COLO Complex 30-day model		
Factor	Parameter Estimate	Variable Coding
Intercept	-3.6601	-
Diabetes	0.0821	Yes= 1 No= 0
ASA Score	0.3028	1= 1 2= 2 3/4/5= 3
Gender	0.1036	Male=1 Female=0
Body Mass Index (BMI)	0.1259	≥ 30= 1 < 30= 0
Patient Age	-0.1396	Patient's age/10
Oncology Hospital	0.5437	Oncology hospital= 1 Non-oncology hospital= 0
Closure	0.2383	Other=1 Primary=0

Risk factors for SSI HYST Complex 30-day model		
Factor	Parameter Estimate	Variable Coding
Intercept	-5.1801	-
Diabetes	0.3247	Yes= 1 No= 0
ASA Score	0.4414	1= 1 2= 2 3= 3 4/5= 4

Body Mass Index (BMI)	0.1106	≥ 30= 1 < 30= 0
Patient Age	-0.1501	Patient's age/10
Oncology Hospital	0.5474	Oncology hospital= 1 Non-oncology hospital= 0

$\log(p/(1-p))=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_iX_i$, where:

p=probability of infection, α=Intercept, β_i=parameter estimate, X_i=value of risk factor (Categorical variables=1 if present, 0 if not present. Refer to 'Variable Coding' column in the table above.

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*)

An expert panel from CDC DHQP was formed to identify potential risk factors in the beginning of model building process. First, all available required clinical relevant variables from NHSN were presented to the expert panel. Patient characteristic variables and the indicator variable for cancer hospitals were considered as potential risk factors. CDC adopted additional risk factors included by American College of Surgeons' National Surgical Quality Improvement Program: diabetes and body mass index.

Also discuss any "ordering" of risk factor inclusion; for example, are social risk factors added after all clinical factors?

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

Stepwise logistic regression model selection methods were used for variable selection. Variables were eligible for entering the model at p-value=0.25 and retaining in the model at p-value=0.05 significant level.

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.** There are no studies providing evidence of a direct relationship between social risk and HAIs. Instead, studies provide evidence that social risk factors are associated with an increased risk of chronic disease conditions, suboptimal care for those conditions, compromised functional status, exposure to nursing homes, and colonization with bacterial pathogens. While these associations may be meaningful they do not establish a direct relationship between social risk and HAIs. Some evidence to the contrary is available. For example, empirical data analyzed by CDC, specifically surgical site infection (SSI) data that hospitals report to NHSN, have yielded findings that SSI risk is lower among older patients compared to younger patients for some surgical procedures: http://www.cdc.gov/nhsn/pdfs/pscmanual/ssi_modelpaper.pdf Until more compelling evidence of a direct relationship between social risk and HAIs is available, it would be premature to adjust for social risk in the clinical quality measures that CDC reports and CMS uses in its Hospital-Acquired Conditions Reduction and Hospital-Value Based Purchasing Programs.

If and when a strategy is warranted for taking social risk into account in the CDC measures, the factors or indices that are included should be demonstrably and statistically associated with increased risk for the HAIs that are measured.

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

Bootstrap sampling method was used to validate the models.

Model validation steps:

1. For each multiple logistic regression model, calculate the c-index as $C_{orginal}$.
2. Generate 100 bootstrap samples from the original dataset with the same number of records as the original sample size using sampling with replacement.
3. For each one of the new samples m=1, …,100, using the predictors of the logistic regression model from step 1 to fit the data with backward elimination approach and calculate the discrimination $as\ C_{boot}^{(m)}$. Note that the model we select from each of the m bootstrap samples could be different from the original model.
4. For each bootstrap sample, the original dataset is used for validation. For this step, the regression coefficients are fixed to their values from step 3 to determine the joint degree of over fitting from both selection and estimation. We obtain $C_{original}^{(m)}$ from this step.
5. For each one of the bootstrap samples, first we will calculate the optimism in the fit: $O^{(m)} = C_{boot}^{(m)} - C_{original}^{(m)}$. Then we obtain O by taking the average of $O^{(m)}$ from M bootstrap samples.
6. The optimism corrected performance of the original model is: $C_{adj} = C_{orginal} - O$. This value is a nearly unbiased estimate of the expected value of the optimism that would be obtained from external validation.

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

If stratified, skip to 2b3.9

Bootstrap validation results for COLO complex 30-day model					
	Estimate	2.5%	97.5%	Number of bootstrap samples	FLAG
Intercept	-3.66	-3.79	-3.53	100	1
diabetes=Y	0.08	0.02	0.14	100	1
ASA score	0.30	0.26	0.34	100	1
gender=Male	0.10	0.06	0.15	100	1
Patient Age10	-0.14	-0.15	-0.13	100	1
BMI>=30	0.13	0.07	0.18	100	1
Closure=Other	0.24	0.14	0.34	100	1
Oncology hospital=Y	0.54	0.36	0.72	100	1

Bootstrap validation results for HYST complex 30-day model						
	Estimate	2.50%	97.50%	Number of bootstrap samples	FLAG	
Intercept	-5.18	-5.40	-4.96	100	1	
diabetes=Y	0.32	0.18	0.47	100	1	

ASA score	0.44	0.37	0.52	100	1	
Patient Age10	-0.15	-0.19	-0.11	100	1	
BMI>=30	0.11	0.03	0.19	100	1	
Oncology hospital=Y	0.55	0.22	0.87	100	1	

Flag=1 means bootstrapped coefficients have the same sign as the original models.

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

c-statistic for COLO complex 30-day model is 0.575.

c-statistic for HYST complex 30-day model is 0.599.

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

Hosmer-Lemeshow p-value for COLO complex 30-day model is 0.64.

Hosmer-Lemeshow p-value for HYST complex 30-day model is 0.012.

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

2b3.9. Results of Risk Stratification Analysis:

N/A because bootstrap method was used.

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*)

With the data reported to NHSN we have made full use of the available risk factor data to produce a series of prediction models for public reporting and pay for performance.

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 SSI measure data are used to calculate an observed/expected ratio, and ratios significantly higher than 1 are indicative of a quality concern that warrants full investigation and response.

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*)

COLO (total of 3493 CCNs – 1888 with SIR calculated):

Significantly greater than 1 = 86 (2% of total, 5% of those w/SIR)

Significantly less than 1 = 112 (3% of total, 6% of those w/SIR)

HYST (total of 3485 CCNs – 746 with SIR calculated):

Significantly greater than 1 = 29 (1% of total, 4% of those w/SIR)

Significantly less than 1 = 27 (1% of total, 4% of those w/SIR)

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?*)

We see variation. We can identify facilities for which summary measure warrants investigation and response.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

Note*: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

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*)

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*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Missing data is not a problem. Business rules are enacted which prevent facilities from entering incomplete records. See also section 2a which provides information about state and CMS validation both of which are motivators for complete data reporting, of which CMS audits can result in financial penalties for identified underreporting.

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*) See above and section 2a

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

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.

Other

If other: Data elements used in the generation of SSI data are routinely generated during care delivery. The NHSN analysis tool will automatically calculate SIRs.

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).

Many of the data fields are available in electronic forms, and facilities may collect the denominator data and report it electronically via HL7 Clinical-Document Architecture file format. Some of the data fields for the numerator are available electronically, from sources such as laboratory reports, while others are found in plain language text.

Some of the data may be available electronically, but not all. SSI remains largely a clinical determination for which free text and structured data will be used for the foreseeable future. In concept, natural language processing (NLP) methods could be brought to bear to convert free text to structured data. However, use of NLP for this purpose would require substantial new work

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

Attachment:

3c. Data Collection Strategy

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

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

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

NHSN

SSI rates and SIR using the methodologies described above have been in use by hospitals participating in CDC surveillance systems since 1986, and the rate measure has been endorsed by NQF in a previous measure set since 2007. Risk-adjusted models for specific operative procedure categories have been developed using aggregate data from over 805 facilities in order to better reflect factors influencing the development of SSI in different patient populations. SIR has proven to be a useful metric for summarizing HAI experience especially when sample sizes within strata are small and when a summary statistic is desired. Business rules built into the software alert users to missing data if no events and/or procedures are reported for a month. The facility is required to confirm that there were no SSIs or COLO or HYST for that month or to enter the missing data. As this measure is tied to Centers for Medicare and Medicaid Services' Value-based Purchasing program, and associated reporting deadlines, the great majority of data is entered within 4.5 months of the end of the quarter.

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)*.

Does not apply—no fees, license, or other requirements.

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)

Quality Improvement (Internal to the specific organization)	Public Reporting Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 Public Health/Disease Surveillance National Healthcare Safety Network http://www.cdc.gov/nhsn/ Payment Program Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html LTCH Quality Reporting Program http://cms.hhs.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/LTCH-Quality-Reporting/index.html Regulatory and Accreditation Programs Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html

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

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

1) Hospital Inpatient Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: Nationwide, currently covers all acute care hospitals with ICUs (approximately 3300).*

Level of measurement and setting: Facility level; acute inpatient hospital
2) Prospective Payment System Exempt Cancer Hospital Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for PPS-Exempt Cancer Hospital to improve health, improve care and lower cost (triple aims)of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: Nationwide, includes all 11 Patient Prospective Payment Exempt Cancer Hospitals in 7 U.S. states with 19,203 average discharges each in FY 2012.*
Level of measurement and setting: Facility level; acute inpatient hospital
3. Hospital Value Based Purchasing
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To reward acute-care hospitals with incentive payments for the quality care provided to Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: Nationwide, includes all 2808 acute care facilities performing these procedures.*
Level of measurement and setting: Facility level; acute inpatient hospital
4. Hospital-acquired Condition(HAC) Reduction Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To provide an incentive for hospitals to reduce HACs
Geographic area and number and percentage of accountable entities and patients: Nationwide, includes all 3216 acute care facilities performing these procedures.*
Level of measurement and setting: Facility level; acute inpatient hospital
*provided by Centers for Medicare and Medicaid Services

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?*)

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.*)

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.

Numerous training materials have been created in order to assist users with the proper understanding and interpretation of this measure. Several webinars and written training materials have been provided. Annual in-person trainings are held to discuss the SIR calculations, risk adjustment, and proper interpretation. Training materials are available online to all hospitals enrolled in NHSN, as well as external partners such as state health departments, quality improvement organizations, and healthcare corporations. NHSN users can run monthly analysis reports within NHSN to view their SIR data. On an annual basis, NHSN publishes national and state-level SIRs in the National and State HAI Progress Report. State health departments that perform validation of SSI data reported to NHSN, provide feedback to facilities.

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.

SIR results are available to NHSN users at any time, based on their current data entry. Data provided within the analysis report includes numerator, denominator, SIR, p-value, and 95% confidence interval. Educational materials are available on the NHSN website that explain each data element. NHSN provides user-support via NHSN@cdc.gov including explanations of data analysis.

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 on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided to us by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection control and surgical associations, and other personnel. An online survey is provided to all live-training attendees who provide feedback on whether objectives were met, usefulness of the training, and whether additional training is needed.

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

Feedback from Hospitals and states: Based on results from a polling survey, hospitals have indicated that they are running SIR analysis reports within NHSN on a monthly basis, and that they use SIRs for prevention activities in their hospital. State health departments are using the SIR for public reporting purposes and to help target facilities for additional prevention. Feedback was received via email regarding the extent of risk adjustment and the limitations.

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

CMS uses this measure for public reporting and payment programs.

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.

Feedback from all stakeholders is considered when developing and implementing the SIR. Different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users. Additional training formats, such as live chats and "quick learn" videos, were created in order to address different training environment that best meet the needs of our audience. We have also provided live demonstrations to users showing how to generate their SIRs in NHSN based on earlier feedback received. If a measure meets the above criteria and they 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.

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.

Please refer to 1b, which outlines the progress in reducing SSIs. SIRs following colon surgeries have been reduced by 6% and SSI SIRs following abdominal hysterectomies by 12% between 2015 and 2016, the most recent years with complete data.

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.

NHSN

Patient medical records and other sources of patient data must be reviewed to determine if the patient meets the necessary criteria for a SSI. It is possible that reviewers may miss symptoms or fail to identify that patients meet criteria thereby underreporting SSI events. Data collectors might also intentionally underreport SSIs. Both of these actions would result in an SIR that is calculated to be lower than actual. Alternatively, patients may be identified as having a SSI when in fact they do not meet SSI criteria and thereby calculate an SIR that is higher than actual. Numbers of operative procedures may be collected inaccurately thereby impacting the SIR. In addition, it is possible SIRs may be miscalculated. The NHSN reporting tool includes business logic to minimize misclassification of SSI.

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

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)

3025 : Ambulatory Breast Procedure Surgical Site Infection (SSI) Outcome Measure

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?

Yes

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

The populations included in the 2 measures differ with the ASC measure being intended for surgeries performed at ambulatory surgery centers and the present measure intended for inpatient surgical patients.

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.)

The populations included in the 2 measures differ with the ASC measure being intended for surgeries performed at ambulatory surgery centers and the present measure intended for inpatient surgical patients. These populations have potential difference in SSI risk as their comorbidities, types of procedures performed, and length of time cared for in a healthcare facility are inherently different. Risk modeling has been performed for both measures, with different models developed based on procedure and facility type. No excess burden collection is anticipated.

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:** Data_Dictionary_SSI___Final_Aug_9-2018-636700284378840296.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Disease Control and Prevention

Co.2 Point of Contact: Daniel, Pollock, dap1@cdc.gov, 404-639-4237-

Co.3 Measure Developer if different from Measure Steward: * Centers for Disease Control and Prevention

Co.4 Point of Contact: Daniel, Pollock, dap1@cdc.gov, 404-639-4237-

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.

None

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: 06, 2015

Ad.4 What is your frequency for review/update of this measure? annuallly, maintenance and adhoc

Ad.5 When is the next scheduled review/update for this measure? 11, 2018

Ad.6 Copyright statement: All CDC documents are public record; no copyright

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments:

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

Brief Measure Information

NQF #: 1716

Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure

Measure Steward: Centers for Disease Control and Prevention

Brief Description of Measure: Standardized infection ratio (SIR) and Adjusted Ranking Metric (ARM) of hospital-onset unique blood source MRSA Laboratory-identified events (LabID events) among all inpatients in the facility

Developer Rationale: The SIR compares a healthcare facility´s performance compared to a national baseline. Facilities are able to see whether the number of LabID events that they have reported compares to the number that would be expected, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.

Numerator Statement: Total number of observed hospital-onset unique blood source MRSA LabID events among all inpatients in the facility per NHSN protocols.

Denominator Statement: Total number of predicted hospital-onset unique blood source MRSA LabID events, calculated from a negative binomial regression model and risk adjusted for facility's number of inpatient days, inpatient community-onset MRSA prevalence rate, average length of patient stay in the hospital, medical school affiliation, facility type, number of critical care beds in the hospital, and outpatient community-onset MRSA prevalence rate from emergency departments and observation units.

Denominator Exclusions: Data from patients who are not assigned to an inpatient bed in an applicable location are excluded from the denominator counts. Denominator counts exclude data from inpatient rehabilitation units and inpatient psychiatric units with different CMS Certification Numbers (CCN) from the acute care facility.

Measure Type: Outcome

Data Source: Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

Level of Analysis: Facility, Other, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Dec 14, 2012 **Most Recent Endorsement Date:** Dec 14, 2012

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 logic model describes how comparing the number of reported (HO) MRSA Bacteremia events to the number predicted drives prevention practices (e.g., appropriate antibiotic use and isolation precautions) that lead to improved outcomes such as reduction in morbidity and mortality associated with MRSA bacteremia.
- The measure is based on the 2006 HICPAC guideline, Management of Multidrug-Resistant Organisms In Healthcare Settings, which provides recommendations for the reduction of transmission of infections within healthcare facilities.
    - 2006 HICPAC guideline included results from over 400 studies.
    - Body of evidence indicates that following the recommended prevention practices can reduce incidence and transmission of MDROs including MRSA in healthcare settings.
    - Additional information provided by the developer on 12/12/18 clarified that (2006) HICPAC edits were made in the February 2017 version indicating text was edited for clarity. The edit does not constitute change to the intent of the recommendations.

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 the evidence provided still applicable and representative of the current state of HO MRSA Bacteremia?*
- *Does the Committee agree with accepting the previous evidence submission?*

Guidance from the Evidence Algorithm

Outcome measure (Box 1) → Relationship between heath outcome and at least one healthcare action is demonstrated by empirical data (Box 2) → Yes→ PASS

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The SIR compares a healthcare facility´s performance compared to a national baseline.
    - National MRSA bacteremia SIR in 2015 is 0.998 = 8,887 observed / 8,906.430 predicted
        - National % change vs. baseline in 2015 < 1%
    - National MRSA bacteremia SIR in 2016 is 0.935 = 8,546 observed / 9,142.247 predicted
        - National % change vs. baseline in 2016 is 6%
    - Percent Change 2016 v. 2015 6% decrease
    - 2015
        - # facilities: 3,616
        - Median: 0.827
        - Range, at 5% and 95%: (0.000 – 2.671)
    - 2016
        - # facilities: 3,602
        - Median: 0.796
        - Range, at 5% and 95%: (0.000 – 2.382)
- Information provided indicates that the national MRSA bacteremia rate as well as facility MRSA bacteremia rates improved from 2015 to 2016.

Disparities

- Patient level social risk factors are not available to be used for risk adjustment or stratification.
- There are no studies showing a direct relationship between social factors and HAIs.
- There are studies showing patients who are found to have had direct or indirect contact with hospitals, care homes or other healthcare facilities have a higher carriage rate than those who are never exposed. Risk for infection is higher in HIV+ patients
- From the literature, the developer notes that among patients hospitalized with acute cardiovascular disease, pneumonia, and major surgery, Asian and Hispanic patients had significantly higher rates of HAIs than white, non-Hispanic patients.

Questions for the Committee:

- Is the information provided enough to demonstrate a continued gap in care?
- Does this gap in care that warrant a national performance measure?
- Has the implementation of this performance measure led to improvements in HO MRSA bacteremia rates nationally and across facilities?

- Is the rationale that social risk factor data are not available and that there is no relationship between social risk and HAIs adequate?

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

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

1a. Evidence
Comments:

**No new evidence was provided b/c the developer stated there is none.

**Pass, an outcomes metric, clinical practice guidelines.

**No concerns.

**As a global report not sure vs incidence reporting.

**Prevention practices can reduce or prevent MDROs.

**There is clearly evidence that these infections are preventable using practices well known in the field and documented by CDC guidances. Unfortunately, CDC chose to highlight a study that is controversial, but it could be the only new evidence since 2012, which is shameful.

**Pass.

**Pass.

**Acceptable.

**Pass.

**Outcome Measure. First endorsed in 2012. No changes in the evidence since last evaluated.

1b. Performance Gap
Comments:

**A performance gap remains. Minimal disparity data was provided.

**moderate, two year data provided showing opportunities for improvement, no disparities studies found linking social factors to HAIs.

**Slow decrease in rates still opportunity to improve, no social factors available.

**Only a slight improvement noted not sure statistically significant.

**SIR range 0-2.382. Disparities show higher incidence in patients with acute CV disease, pneumoniz, major surgery, Asian and Hispanic from the literature. Disparities not measured.

**Definitely there is a performance gap among hospitals that warrants this measure nationally. The question is whether this measure articulates that gap clearly - to the public and to hospitals. NHSN states that 8% of hospitals have significantly higher infection ranking than the baseline and 4% rank lower (the desired rank). That leaves 88% of the hospitals in the HUGE middle called "no different than the predicted SIR" - I think it would help consumers and hospitals if that middle group was broken up, using at least 5 increments of scoring; also some confidence intervals seem quite wide. The measure is an outcome measure, which is good. It has demonstrated improvements between 2015-2016, but CDC should have showed us improvements since the last endorsement in 2012. They cannot compare with the current baseline set in 2015, but they could show improvements from 2012-2015 in addition to 2015-2016 - they have the data. It would also be good for them to articulate how the national baseline changed in 2015 - that should be further evidence of improvement. Disparities were not measured, mostly because hospitals do not collect sociodemographic information and they are the source for NHSN data. However, I believe CDC could do a better job at identifying whether people of color or with low incomes have more infections - especially important with MRSA infections as several studies have identified higher rates in black men. Also,

one of the risk adjustment factors is whether the hospital is a medical school - begs the question why? is it because patients are being cared for by physicians in training or whether the population served by these hospitals are typically low income with limited access to health care. Regardless, a truer picture of the performance of these facilities might be available if the infection calculations connected to medical schools were not modified. Comparing the medical schools with each other would be a fairer way to see performance gaps among similar hospitals.

**Reduction between '15 and '16 reported.

**Moderate.

**Modest improvement 12016-2016; appears there are some really poor performing facilities.

**Still room for improvement.

**Information provided indicates that the national MRSA bacteremia rate as well as facility MRSA bacteremia rates improved from 2015 to 2016. There are studies showing patients who are found to have had direct or indirect contact with hospitals, care homes or other healthcare facilities have a higher carriage rate than those who are never exposed. Risk for infection is higher in HIV+ patients • From the literature, the developer notes that among patients hospitalized with acute cardiovascular disease, pneumonia, and major surgery, Asian and Hispanic patients had significantly higher rates of HAIs than white, non-Hispanic patients.

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.

Composite measures only:

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Evaluation of Reliability and Validity (and composite construction, if applicable):

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0; M-5; L-0; I-0
- Validity: H-0; M-4; L-1; I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability testing considered to be met because data element validation were conducted.
- NQF criteria states that additional reliability testing is not needed if empirical validity testing of patient-level data is conducted and the results are adequate.
- Panel member noted that while testing information meets minimum requirements, they would have liked to see a separate reliability testing of data elements.

Validity

- Validity testing was performed for data element.
- Data Element
    - Validation using a sample of charts within a sample of facilities within 5 states in varying years reviewed by trained chart abstractors and tallied against data reported to the National Healthcare Safety Network (NHSN).
    - Case classification during the medical chart review and application of the protocol by the auditor is considered as the gold standard and compared with the facility determinations.
    - Developer provides sensitivity/specificity/PPV/NPV seemingly for the MRSA variable only. However, panel members noted that testing of variables included in the risk adjusted model was not reported; no information is provided on the validity of data elements used for risk adjustment and to identify the denominator population.
    - Results:

		Sensitivity	**Specificity**	**Positive predictive value**	**Negative predictive value**
Tennessee	2015	80.9%	87.5%	97.5%	42.8%
Wisconsin	2009	95.2%	63.6%	93.7%	70%
New Mexico	2016	98.7%	100%	100%	98.8%
California (MRSA/VRE BSI)	2014	88%	NP*	NP*	NP*
Maine	2015	83%	NP*	74%	NP*

*NP- Not provided

- Risk Adjustment

- This is a risk-adjusted model with 6 risk factors: Inpatient community onset prevalence; average length of stay, medical school affiliation; facility type; number of ICU beds; and outpatient community onset prevalence.
- No social risk factors were included because these were not collected in the NHSN for all patients in the patient population.
- The risk model was conducted using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. The multivariate regression model was confirmed and validated using bootstrap validation techniques.
- Results:
  - The p-values for all variables in the final multivariate model were statistically significant, with several variables having a p-value < 0.0001.

Standing Committee Action Item(s):

- The Standing Committee can discuss the reliability and validity testing, or agree to take the ratings of the Scientific Methods Panel.

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)?
- The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

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

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

Evaluation A: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 1716

Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital- onset Methicillin-resistant Staphylococcus aureus bloodstream infection(MRSA) Bacteremia Outcome Measure

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

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

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☒ **Population: Regional and State**
☐ **Integrated Delivery System** ☒ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

   PANEL MEMBER 2: Yes, the numerator is the number of observed hospital-onset unique blood source MRSA LabID events among all inpatients in the facility per NHSN protocols and the denominator is the number of predicted such events calculated from an adjusted model. This measure has been adopted nationally and applied to facilities for at least ten years.

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

   PANEL MEMBER 1: The measure requires lab confirmed MRSA bacteremia Lab ID events and appears to be part of a larger series of Standardized infection ratios monitored for public health and quality improvement purposes. The measure assumes all significant events result in a blood test with either positive or negative findings. The authors do not address situations where the infection may be present but no test ordered or errors in the lab's handling of samples – the underlying assumption is that this is a serious problem always resulting in testing. Auditing appears to focus on the accuracy of reporting. It would be helpful to know that 'errors of omission' (suspected cases with no test or lab errors resulting in no test or the wrong test) are not a problem here.

   PANEL MEMBER 2: I have no concerns about the measure specs.

   PANEL MEMBER 3: none

   PANEL MEMBER 4: None.

   PANEL MEMBER 5: Sensitivity generally high, specificity and negative predictive value vary substantially across states. Probably adequate, but would like to see discussion of efforts to improve accuracy of reporting.

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**

   PANEL MEMBER 1: The unit of analysis for this measure is the hospital rate of infection. This rate seems to then be aggregated to the unit of interest, but that a State or some other unit. Results are presented for hospitals and states.

   PANEL MEMBER 2: Previously tested, not re-tested with this submission

   PANEL MEMBER 3: No testing was conducted

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**

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

   Submission document: Testing attachment, section 2a2.2

PANEL MEMBER 1: The authors use trained auditors to compare rates in the medical record against those reported by the hospital to NHSN. This seems appropriate to determine the reliability of the hospital's own reporting.

PANEL MEMBER 2: What was provided was data regarding the sensitivity, specificity, sensitivity, positive predictive value, and negative predictive values across five states which were felt to be supportive of the measure's reliability.

PANEL MEMBER 4: The assertion is that "widespread" was insufficient.

PANEL MEMBER 5: The basic construction of the measure is:

Get count of lab-based measures of MRSA. Divide this by patient days to get rate. Estimate expected rate from negative binomial regression model. Construct actual to expected by dividing actual rate to estimated rate.

Potential sources of unreliability of measure:

a. Errors in counts of events.
b. Variability in counts over time due to random fluctuation
c. Errors in counts of patient day denominator.
d. Imprecision in risk adjustment model

Methods used:

a. Reabstracting and assessment of sample of charts and calculation of specificity, sensitivity, positive predictive value, and negative predictive value. Method is appropriate.
b. Not done. Given low counts (0-25, IQR 0-1) , failure to consider year to year variability due to randomness is a weakness.
c. Not done, but this should be a de minumus source of error.

Variables in regression model initially identified through expert panels based on available data taking burden into account. Continuous variables used to construct categorical variables (community infection rate dichotomized; average LOS terciles; number of ICU beds quintiles). Other categorical variables (hospital type: cancer, general acute, other specialty; medical school affiliation: major, grad/undergrad, none). Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but values not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on "optimism in the fit." No assessment made of stability of SIRs across bootstraps or change in ranking across significantly higher, significantly lower, although only 8% in higher and 4% in lower category

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

   Submission document: Testing attachment, section 2a2.3

PANEL MEMBER 1: They do not report detailed results of the auditing, but do not report any concerns with the results.

PANEL MEMBER 2: Reliability testing for critical data elements was not provided with this submission. "No additional testing was conducted as the value of the measure as an indicator for differentiating good and poor

performance has been substantiated by its broad use for that purpose. The measure is widely used by healthcare facilities and state health departments".

PANEL MEMBER 3: Developer did not submit reliability testing

PANEL MEMBER 4: N/A.

PANEL MEMBER 5: Sensitivity, specificity, positive predictive, negative predictive value vary from year to year and state to state. Probably meet minimum standards but would like to see discussion with developer of efforts to improve reporting over time.

a. Given low counts (0-25, IQR 0-1) , failure to consider year to year variability due to randomness is a weakness.
b. NA
c. The developers do not report the "optimism" estimate or any other data to assess the stability of the SIRs or rankings over the bootstrapped models. The 2.5-97.5% CIs are wide for coefficients on some variables (e.g., 0.269-0.1645 for the middle LOS category, indicating a range for multiplying the rate from 3% to 18%). The risk model itself allows for a 4-fold increase in the estimated rate within the General Acute Hospital category between the reference category and highest category for all categorical variables, so there can be substantial shifts in estimated/expected depending on the risk model, with the sensitivity of the estimates or rankings to variations in estimated coefficients not presented.

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 <u>all</u> testing results):

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

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

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

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

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: The measure developers could have presented some or the statistics on hospital versus auditor reported infection rates. They also could have looked at reliability within facilities over time and included more states or facilities in their analysis of reliability. Undoubtedly this would be expensive, but it seems like a worthwhile exercise for this important public health measure.

PANEL MEMBER 2: No additional testing performed with this submission. Reference made to widespread utilization of the measure from initial endorsement.

PANEL MEMBER 3: Developer states that testing is not necessary because the measure is widely used. I do not think that is a sufficient reason to not test for reliability.

PANEL MEMBER 4: None provided.

PANEL MEMBER 5: There is inadequate information presented on the stability of the measure to reasonable expected variation in lab reported counts of MRSA or the precision and stability of the risk adjustment model.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

PANEL MEMBER 1: I was initially concerned about the exclusion of quarters with zero patient days, but in the testing sample, this only resulted in dropping 295 quarters out of a total of 14,132 quarters. That said, it seems like this is valid information that should be included in the measure.

PANEL MEMBER 2: I have no concerns with exclusions. Basically, if there were missing or zero denominators.

PANEL MEMBER 3: No testing of the potential impact of missing data or the reason that data might be missing (root cause).

PANEL MEMBER 4: None.

PANEL MEMBER 5: None

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

Submission document: Testing attachment, section 2b4.

PANEL MEMBER 1: The authors point out that the measure varies between facilities cross-sectionally and over time. They feel the differences are meaningful and suggest the wide use of the measure is an indication of its value to providers and government policy makers. More data on the within versus between facility variation in the measure would have been helpful. Similarly, information on variation by state over time would help create a more complete picture.

PANEL MEMBER 2: Citation is given that the median SIR score is 0.827 and the national pooled mean in 0.998 amongst 3,616 facilities. There are 8% higher than the national SIR and 4% lower than the national SIR.

PANEL MEMBER 3: Level of significance not reported in 2b4.2 – not clear what 'significantly higher/lower' really means without that statistic.

PANEL MEMBER 4: Describe sensitivity and specificity as "reliability" but these are assessments of validity. There is variation at the state level.

PANEL MEMBER 5: See discussion re reliability in items 7a,d, 11.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

Submission document: Testing attachment, section 2b5.

PANEL MEMBER 1: I do not believe comparability is an issue here – there is only one set of specifications.

PANEL MEMBER 2: No concerns as the data sources are the same.

PANEL MEMBER 4: None.

PANEL MEMBER 5: NA

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

Submission document: Testing attachment, section 2b6.

PANEL MEMBER 1: No major concerns about missing data for the measure itself.
PANEL MEMBER 2: Pre-emptive alerts are given for , and if not corrected, facility does not receive a SIR
PANEL MEMBER 4: None.
PANEL MEMBER 5: NA

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☒ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☒ No

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

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

PANEL MEMBER 1: I assume that the model development and validation work used two different data sets – this is not clear from the write up, but it important given the bootstrapping and stepwise selection used to generate the models. I also assume the covariates in the model are for the time period before the measurement period. In other words, if the predicted rate is for 2015 then the covariates are from 2014. This is important because many of the covariates are in fact MRSA rates at the facility or unit level. Given these assumptions, the model has good face validity, is parsimonious and good statistical properties.

PANEL MEMBER 2: The risk adjustment model was conducted using a negative binomial regression. Univariate models were first constructed to assess relationships between the risk factor and the MRSA incidence rate, then applied to a multivariate model. Selection criteria were eligibility for inclusion at a p value of 0.25 and retention at a p value of 0.05. In the multivariate model, forward selection was utilized based on the lowest Wald Chi-square value. Goodness of fit was applied at each modeling step using the AIC statistics. The final model was then confirmed using backwards elimination, starting with the highest p value. Model validation was tested by a bootstrap sampling method and the results are provided. The values for all variables in the final multivariate model were statistically significant, with several less than 0.0001.

Potential risk factors were selected based on availability in the source database, NHSN, literature review, and subject matter expert opinion. Social risk factors were not specifically included due to data entry burden and a cited lack of evidence that supports the hypothesis that data collection of such would justify inclusion.

PANEL MEMBER 3: Risk adjustment approach is appropriate

PANEL MEMBER 5: Variables in regression model initially identified through expert panels based on available data taking burden into account. Continuous variables used to construct categorical variables (community infection rate dichotomized; average LOS terciles; number of ICU beds quintiles). Other categorical variables (hospital type: cancer, general acute, other specialty; medical school affiliation: major, grad/undergrad, none). Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but values not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on "optimism in the fit." No assessment made of stability of SIRs across bootstraps or change in ranking across significantly higher, significantly lower, although only 8% in higher and 4% in lower category.

- The developers do not report the "optimism" estimate or any other data to assess the stability of the SIRs or rankings over the bootstrapped models. The 2.5-97.5% CIs are wide for coefficients on some variables (e.g., 0.269-0.1645 for the middle LOS category, indicating a range for multiplying the rate from 3% to 18%). The risk model itself allows for a 4-fold increase in the estimated rate within the General Acute Hospital category between the reference category and highest category for all categorical variables, so there can be substantial shifts in estimated/expected depending on the risk model, with the sensitivity of the estimates or rankings to variations in estimated coefficients not presented.
- I'm willing to accept the variables included in the model as a potentially reasonable basis for differentiating expected performance across hospitals. The community acquired rates seem like a useful adjuster. There is, however, inadequate information presented on the precision and stability of the risk adjustment model

VALIDITY: TESTING

17. **Validity testing level: ☒ Measure score ☒ Data element ☐ Both**
18. **Method of establishing validity of the measure score:**
    - ☐ **Face validity**
    - ☒ **Empirical validity testing of the measure score**
    - ☒ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    - **PANEL MEMBER 1**: The authors are looking at sensitivity and specificity cross-setionally, comparing rates across states. They did not look at consistency within facilities or states over time. They also look at mean and median rates for their national sample of hospitals, suggesting differences between these two numbers indicate variation in the measure. They do not report results for other levels of analysis.
    - **PANEL MEMBER 2**: What was provided was data regarding the sensitivity, specificity, positive predictive value, and negative predictive values across five states which were felt to be supportive of the measure's reliability.
    - **PANEL MEMBER 3**: Validity testing in 5 states from 2009 to 2016
    - **PANEL MEMBER 4:** Adequate, although there is variation by state.
    - **PANEL MEMBER 5**: Chart review of numerator.
20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    - **PANEL MEMBER 1:** The authors present results for 5 states with 2 state having incomplete information (non-reporting of sensitivity, PPV and NPV). One state has a rather low Negative Predictive Value, but the authors note the range for sensitivity and specificity has been smaller and higher in their most

recent data (2014-2016) (sensitivity range from 81-99% and specificity range of 87-100%). This sounds very good for a wide range of different hospital types (e.g., community based, academic and so on).

- **PANEL MEMBER 2:** The process and data provided in Q16 above is used to establish validity of the model to derive the SIR score.
- **PANEL MEMBER 3:** Difficult to assess – there is not sample size listed for the states nor is there any statistical testing provided.
- **PANEL MEMBER 4:** See 19.
- **PANEL MEMBER 5:** Estimates of sensitivity, specificity, positive and negative predicted value from chart review of sample.

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☒ **No**

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

    PANEL MEMBER 1: The authors feel that wide use of the measure indicates its validity. They do not present any other information on high or low performers. It would have been helpful, for example, to include the scores for other infection measures. One could hypothesize that having a poor score on one infection measure would correlate with other infection measures.

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

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

    ☒ **Yes**

    ☐ **No**

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

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

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

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

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

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

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

    PANEL MEMBER 1: The authors are primarily focused on re-estimating their risk model and presenting these results. There is much less information on the data they have collected over time, the sample they are using for validation and external indicators of infection rates above and beyond the hospital electronic and paper health records. The measure appears to focus on hospital, but the developers see the measure as a good indicator of population health. More analysis at the population or group level would be helpful. Given the limits of what is presented here, the measure seems to warrant a low to moderate overall validity rating.

PANEL MEMBER 2: A formal testing of the score is not provided. What is provided in the rationale for model development and data showing national and regional differences.

PANEL MEMBER 4: Sensitivity, specificity, PPV, NPV data provided appear adequate.

PANEL MEMBER 5: Ideally, sensitivity, specificity, positive and negative predictive values would be higher, particularly specificity and negative predictive value, but within range of acceptability

ADDITIONAL RECOMMENDATIONS

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

PANEL MEMBER 2: This was a very difficult measure to assess. It has strong national usage and application but the document provided does not provide either interval information regarding testing or the details needed to make thorough assessment. Much is deferred to a prior endorsement. The steering committee assessment of 2012 is provided and is noted to be high on importance, moderate+ on reliability and validity, high on usability and high on feasibility. The measure does provide extensive detail on the model's development for risk stratification.

PANEL MEMBER 5: The standing committee should discuss the basis for expert panel assessment of variables to be included in risk adjustment and extent to which variations in results across potential risk adjustment models should be presented.

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

2a1. Reliability – Specifications
Comments:

**No concerns.

**Specifications provided.

**Overall risk adjustment and reliability passes-would like to hear discussion between scientific group and CDC for some questions raised but overall measure acceptable to me.

**How is colonization accounted for?

**Measures clearly stated. Risk for underreporting especially if testing not done.

**The specifications for the data elements required of hospitals are clear, but the problems lie with the execution. Limited validation is conducted across the US, mostly due to funding issues. There have always been concerns that some hospitals are gaming the system, although it is probably not widespread. More regular validation activities might identify this. Numerous states do validation, as it indicated - I believe more states than are presented here (pretty sure WA and NY validate their data).

**Agree with prelim moderate rating.

**Moderate.

**Acceptable.

**No concerns.

**Reliability testing for critical data elements was not provided with this submission. "No additional testing was conducted as the value of the measure as an indicator for differentiating good and poor performance has been substantiated by its broad use for that purpose. The measure is widely used by healthcare facilities and state health departments". Developer did not submit reliability testing.

2a2. Reliability – Testing
Comments:

**No concerns.

**Moderate, data element validation conducted.

**No.

**No.

**Sensitivity, specificity, PPV and NPV over 5 states, auditors to compare rates in facilities with those reported to NHSN.

**I do not have concerns with the variation in lab reported cases. This is really the only hard evidence regarding hospital onset MRSA bacteremia. CDC could do spot checks to determine if tests were being avoided, but my guess is that the lab tests are typically returned after the patient leaves. I personally know of some of these cases. I do have concerns that this measure identifies only the tip of the iceberg regarding serious MRSA infections. Unfortunately, our health care system is not equipped to do follow up with patients in a meaningful way. Most of these infections are not hospital onset, but the symptoms arise AFTER the patient is discharged. If a patient then returns to the hospital with HAI symptoms, it is unclear whether they are counted as community acquired and mixed in with the community rate of MRSA used for risk adjustment or if they are counted as hospital acquired.

**No.

**No.

**None.

**By now steward ought be able to test reliability across a much broader swath...not just 5 states.

**Acceptable.

**Not submitted for this submission. There is inadequate information presented on the stability of the measure to reasonable expected variation in lab reported counts of MRSA or the precision and stability of the risk adjustment model.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences

Comments:

**No concerns.

**Moderate, results provided.

**No.

**No.

**Reported sensitivity and specificity across 5 states. 2 states incomplete though ranged from 81% to 100%.

**Yes - I have concerns with the risk adjustment. I agree with the panel that CDC should provide the data behind the risk adjustment and those should be validated in some way. Of particular concern is the rate of MRSA colonization in the community's general population. I believe these are calculated by the hospital and should be validated. I don't believe they explained adequately why inpatient days are included in determining the denominator.

**No.

**Moderate.

**Appreciate SM commenter's suggestion about correlations between this measure and other HAIs; would like to see that.

**Acceptable.

**The authors feel that wide use of the measure indicates its validity. They do not present any other information on high or low performers. It would have been helpful, for example, to include the scores for other infection measures. One could hypothesize that having a poor score on one infection measure would correlate with other infection measures.

**No concerns.

**No concerns.

**No.

**No.

**Would like to see patient level social risk factors.

**I do not believe missing data is a threat to the validity of the measure. I am concerned that some of the efforts to make performance "comparable" are unexplained - there is no information about the reasoning behind the 6 types of risk adjustment, beyond stating that a group of experts came up with these.

**No.

**None.

**Some sites clearly are really poor performers.

**No concerns.

**This was a very difficult measure to assess. It has strong national usage and application but the document provided does not provide either interval information regarding testing or the details needed to make thorough assessment. Much is deferred to a prior endorsement.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**No concerns about exclusions.

**Risk adjusted model with 6 risk factors.

**Risk adjustment appears adequate.

**Is the risk model capturing the entire hospital picture of risk?

**Exclusions clearly defined and appropriate.

**We were not provided information about whether the risk adjustment was appropriately developed or tested or validated. They say they did statistical tests, but they are not provided to the committee. It seems to be quite complicated and I have concerns that they may hide poor performing hospitals through adjustments. The sources of the information for risk adjustment only described in a footnote on page 34 of the worksheet. An annual hospital survey is the source of much of the data, but it is not stated whether this is an NHSN survey or from some other source. Especially troubling is the adjustment for high levels of colonization in the community served by a hospital, the source of which is not clearly articulated. Making such an adjustment seems to indicate that infections cannot be prevented in patients colonized with MRSA - which is untrue. Why should a hospital in a community with high colonization rates not be expected to prevent these very dangerous serious invasive infections? Back to how community colonization is determined, one real predictive factor of colonized patients (which is mentioned in the worksheet) is patients who have had prior contact with the health care system - "There are studies showing patients who are found to have had direct or indirect contact with hospitals, care homes or other healthcare facilities have a higher carriage rate than those who are never exposed." I wonder if this would be a more accurate factor to use. Most states that require high risk incoming patients to be screened use this as a factor. I believe the V.A. screens all incoming and outgoing patients for MRSA colonization and that system has significantly reduced MRSA infections.

**None – moderate.

**Appreciate the limitations of NHSN re: social determinants.

**Appropriate.

**All data followed the same exclusion rules: quarter with zero patient days, or hospitals with missing survey variables (i.e. risk-adjustment variables) were excluded.

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 data elements are in defined fields in electronic sources.
- NHSN provides the option for facilities to collect the data electronically and download into NHSN. They leave the option for manual entry for facilities that are not equipped or ready to submit electronically.

Questions for the Committee:

- Are there any difficulties the Committee is aware of regarding the feasibility of the measure?
- During the previous review, the Committee was concerned that lab tests confirming MRSA may not ordered by hospitals in order to artificially reduce the number of MRSA infections reported. Is there any indication of this type of situation and is this still a concern?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**No concerns about feasibility.

**High, data collected through medical records or NHSN electronic form.

**Data can be either electronic or paper-no concerns.

**Okay.

**Defined in electronic sources. NHSN tool.

**The collection of this data is feasible and NHSN allows hospital to report in numerous ways. The agency has also worked with IT companies to make sure the EHRs have the correct elements for e-reporting.

**Agree with high feasibility prelim rating.

**High lab based.

**No concerns.

**No concerns.

**The NHSN Multidrug Resistant Organism and C. difficile Infection (MDRO/CDI) module has been available for facilities to use since 2009. The ability to perform facility-wide surveillance with a single denominator was introduced in 2010, reducing data collection burden on participating facilities.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

- The measure is used in numerous public reporting and payment programs: Hospital Inpatient Quality Reporting Program (HIQR), Prospective Payment System Exempt Cancer Hospital Quality Reporting Program, Inpatient Rehabilitation Facility (IRF) Quality Reporting Program, Long Term Care Hospital (LTCH) Quality Reporting Program, Hospital Value-Based Purchasing, and Hospital-Acquired Condition Reduction Program (HACRP).
- The measure is also used for Public Health/Disease Surveillance.

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

- Feedback to the developer is provided by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection prevention and other personnel.
    - Based on results from a polling survey, hospitals have indicated that they are running SIR analysis reports within NHSN on a monthly basis, and that they use SIRs for prevention activities in their hospital.
    - State health departments are using the SIR for public reporting purposes and to help target facilities for additional prevention. Feedback was received regarding the extent of risk adjustment and the limitation.
- In response to feedback, different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users and additional training formats and demonstrations were created to assist with understanding, interpreting, and implementing this measure.
- This measure meets all three criteria for feedback, although additional details regarding the feedback received (number of measured entities given measure results/assistance with implementation or interpretation, number of entities that provided feedback) would be helpful.

Additional Feedback: N/A

Questions for the Committee:

- Has the measure been appropriately vetted in real-world settings by those being measured?

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

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

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

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

Improvement results

- There has been major progress since 2005 in preventing MRSA bacteremia due to declines in hospital-onset and community-onset, healthcare-associated bacteremia. However, declines are slowing in these areas.
- Using the 2015 baseline, there was a 5% decline in SIR between 2015 and 2016.
- There was a slow continuous decline in the unadjusted NHSN crude rate of hospital-onset MRSA bacteremia (the outcome represented by the SIR) from 2012 through 2016, ranging from 0.61 cases per 10,000 patient days to 0.55 cases per 10,000 patient days, with no increase in 2015.

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

N/A

Potential harms

- It is possible that medical record reviewers will miss positive cultures or important dates that would indicate that a LabID event should be recorded.
- Reviewers might miss data in the medical record that would indicate a positive culture should not result in a LabID event.
- It is possible that data abstractors could intentionally underreport LabID events.
- Business logic is built into the NHSN application to minimize incorrect entry of LabID events.
- Agencies have indicated interest in performing validation of LabID event surveillance.

Additional Feedback: N/A

Questions for the Committee:

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

RATIONALE:

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

4a1. Use - Accountability and Transparency
Comments:
**Already in use in accountability programs.
**High, publically reported by many organizations including CMS, NHSN runs analysis, hospitals use SIR analysis for prevention activities.
**Publically reported not clear but progress is slow.
**General results are not applicable but cluster outbreaks are.
**Being used in HIQR, Cancer hospital quality reporting, LTCHquality reporting, hospital value based purchasing, HAC. Feedback is that it is being used to target prevention activities. Concerns about risk adjustment and limitations.
**Yes - hospitals have access to their data regularly and many of them use the data for quality improvement. Also, many health departments use the results to target hospitals who may need training or inspecting. I am aware that over the years, many users have provided CDC with feedback on their use of these measures. Hopefully, that has resulted in better performance measurement, but also could weaken the measure if NHSN is too accommodating.
**Agree with prelim rating.
**Pass.
**Used in payment programs.
**No concerns.
**SIR results are available to NHSN users at any time, based on their current data entry. Data provided within the analysis report includes numerator, denominator, SIR, p-value, and 95% confidence interval. Educational materials are available on the NHSN website that explain each data element. NHSN provides user-support via NHSN@cdc.gov including explanations of data analysis.

4b1. Usability – Improvement
Comments:
** No concerns.
** High, some risk of underreporting and possibly some missed data.
** Performance can be used to improve performance.
** I think this needs to be linked to a measure of colonization.
** Address social risk factors.
** This measure has demonstrated usability for hospitals and as a consumer advocate, I think it is an important measure. It would be more useable if the reporting was done with more increments rather than simply bunching 88% of the hospitals as "no different from the baseline." While some states/hospital are working to get to zero, most are probably not concerned unless they are in the outlier "significantly higher" group. Average/middle ground is not the goal and the way this measure (and other HAI measures) are provided to the public fails to help us discern which hospitals in that big "middle" are closer to becoming outliers.
**Minimal apparent unintended consequences.
**None.
**No concerns.
**None identified.

Criterion 5: Related and Competing Measures

Related or competing measures
Related
1717: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure

Harmonization

These measures appear to be harmonized to the extent possible.

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

5. Related and Competing

Comments

**None.

**None.

**No.

**No.

**Harmonized.

** I know of none and the developer reported none. It would be great for CDC to develop a broader measure that brings in additional MRSA infections, which are a big problem in our hospitals and communities (most of which originally began from people colonized while in health care facilities).

**Related to 1717 – CDI.

**Already harmonized.

** No concerns.

** Related measure but not competing - different organism.

** 1717: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure. These measures appear to be harmonized to the extent possible.

Public and Member Comments

No NQF members have submitted support/non-support choices as of: 01/22/2019

Public Comment

** The Federation of American Hospitals (FAH) appreciates the opportunity to comment on NQF #1716: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus bloodstream infection. FAH requests that the Patient Safety Standing Committee consider whether sufficient information has been provided regarding the data element validity testing under Criterion 2b. Validity. The measure developer notes that the validation was completed on a sample of hospitals and patient charts in each state but we were unable to determine whether the sampling was sufficient and question whether the information aggregated at the state level rather than for each facility and at the measure score and not for each individual data element demonstrates valid data capture and reporting at the facility level. We believe that additional information to demonstrate the validity of each data element by facility is needed to meet the validity criterion.

Support/Non-Support

- No NQF Members have submitted support/non-support choices as of this date.

Brief Measure Information

NQF #: 1716

Corresponding Measures:

De.2. Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure

Co.1.1. Measure Steward: Centers for Disease Control and Prevention

De.3. Brief Description of Measure: Standardized infection ratio (SIR) and Adjusted Ranking Metric (ARM)of hospital-onset unique blood source MRSA Laboratory-identified events (LabID events) among all inpatients in the facility

1b.1. Developer Rationale: The SIR compares a healthcare facility´s performance compared to a national baseline. Facilities are able to see whether the number of LabID events that they have reported compares to the number that would be expected, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.

S.4. Numerator Statement: Total number of observed hospital-onset unique blood source MRSA LabID events among all inpatients in the facility per NHSN protocols.

S.6. Denominator Statement: Total number of predicted hospital-onset unique blood source MRSA LabID events, calculated from a negative binomial regression model and risk adjusted for facility's number of inpatient days, inpatient community-onset MRSA prevalence rate, average length of patient stay in the hospital, medical school affiliation, facility type, number of critical care beds in the hospital, and outpatient community-onset MRSA prevalence rate from emergency departments and observation units.

S.8. Denominator Exclusions: Data from patients who are not assigned to an inpatient bed in an applicable location are excluded from the denominator counts. Denominator counts exclude data from inpatient rehabilitation units and inpatient psychiatric units with different CMS Certification Numbers (CCN) from the acute care facility.

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

S.20. Level of Analysis: Facility, Other, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Dec 14, 2012 **Most Recent Endorsement Date:** Dec 14, 2012

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?

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

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

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

NQF_evidence_attachment__Final_July_27-636683017561311681.docx,1716_Evidence_MSF5.0_Data-v5.doc

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*)**:** 1716

Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Healthcare facility Onset (HO) Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 7/27/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.

- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☒ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

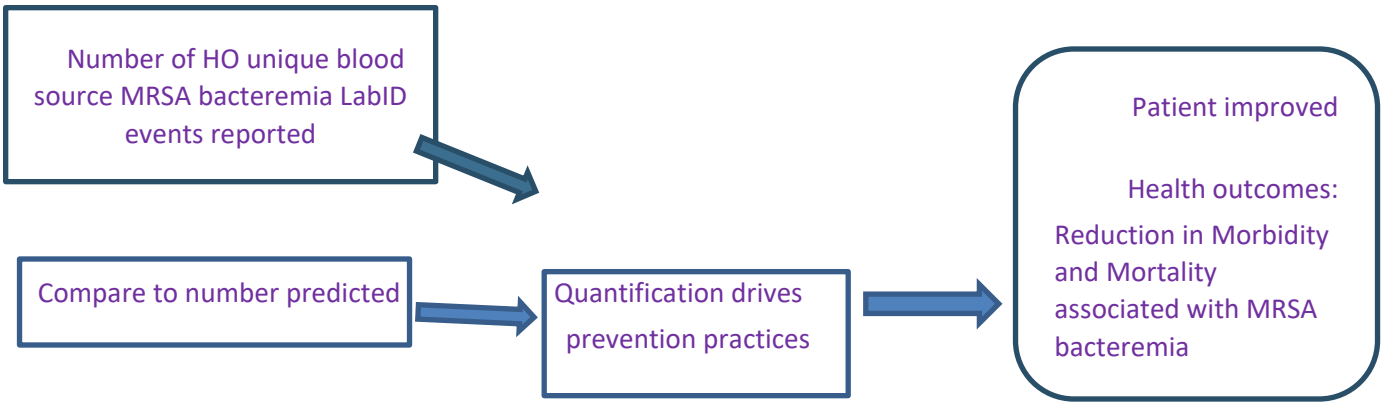
☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

The SIR describes a healthcare facility´s performance compared to a national baseline. Facilities are able to see how the number of (HO) MRSA Bacteremia events they have reported compares to the number predicted, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.



*Measurement leads to appropriate antibiotic use, isolation precautions, HO incidence rate decline

Clinical practice guidelines for the management of multidrug resistant organisms, including MRSA, have been published. Adherence to the recommendations in the guidelines can result in decreased rates of MDRO transmission and infection. Decreasing rates of infection will result in a lower SIR, which indicates improving performance.

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

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

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

A wide ranging variety of studies examining hospital-onset MRSA bacteremia infection rates and process measures exist. In 2006, the Healthcare Infecton Control Practices and Advisory Committee (HICPAC) published a clinical guideline for managing MDROs in the healthcare setting, which is where this measure is focused.

The 2006 HICPAC guideline, Management of Multidrug-Resistant Organisms In Healthcare Settings, included results from over 400 studies. The 2006 HICPAC guideline for management of MDROs in healthcare settings provides recommendations for the reduction of transmission of infections within healthcare facilities. As is standard with all HICPAC guidelines, recommendations were categorized on the basis of existing scientific data, theoretical rationale, applicability, and economic impact. The recommendations in the 2006 HICPAC guideline can consistently be used to reduce the incidence and transmission of infections with MDROs in healthcare facilities. If there is contradictory evidence of efficacy of a prevention practice, a recommendation is not made. The body of evidence reviewed in the preparation of the 2006 HICPAC guideline indicates that following the recommended prevention practices can reduce incidence and transmission of MDROs including MRSA in healthcare settings.

A patient outcome example is provided: CHG bath

https://www.ncbi.nlm.nih.gov/pubmed/25274761

BACKGROUND:

Methicillin-resistant Staphylococcus aureus (MRSA) is a virulent organism causing substantial morbidity and mortality in intensive care units. Chlorhexidine gluconate, a topical antiseptic solution, is effective against a wide spectrum of gram-positive and gram-negative bacteria, including MRSA. Objectives To examine the impact of a bathing protocol using chlorhexidine gluconate and bath basin management on MRSA acquisition

in 5 adult intensive care units and to examine the cost differences between chlorhexidine bathing by using the bath-basin method versus using prepackaged chlorhexidine-impregnated washcloths.

METHODS:

The protocol used a 4-oz bottle of 4% chlorhexidine gluconate soap in a bath basin of warm water. Patients in 3 intensive care units underwent active surveillance for MRSA acquisition; patients in 2 other units were monitored for a new positive culture for MRSA at any site 48 hours after admission.

RESULTS:

Before the protocol, 132 patients acquired MRSA in 34333 patient days (rate ratio, 3.84). Afterwards, 109 patients acquired MRSA in 41376 patient days (rate ratio, 2.63). The rate ratio difference is 1.46 (95% CI, 1.12-1.90; P = .003). The chlorhexidine soap and bath basin method cost $3.18 as compared with $5.52 for chlorhexidine-impregnated wipes (74% higher).

CONCLUSIONS:

The chlorhexidine bathing protocol is easy to implement, cost-effective, and led to decreased unit-acquired MRSA rates in a variety of adult intensive care units.

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

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

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	

Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

1a.4.3. Provide the citation(s) for the 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 SIR compares a healthcare facility´s performance compared to a national baseline. Facilities are able to see whether the number of LabID events that they have reported compares to the number that would be expected, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.

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.

When SIRs are compared over time, assessment of performance can be made. CDC has demonstrated significant performance gaps in SIRs across facilities. See below:

National MRSA bacteremia SIR in 2015 is 0.998 = 8,887 observed / 8,906.430 predicted

National % change vs. baseline in 2015 < 1%

National MRSA bacteremia SIR in 2016 is 0.935 = 8,546 observed / 9,142.247 predicted

National % change vs. baseline in 2016 is 6%

Percent Change 2016 v. 2015 6% decrease

2015-

facilities: 3,616

Median: 0.827

Range, at 5% and 95%: (0.000 – 2.671)

2016-

facilities: 3,602

Median: 0.796

Range, at 5% and 95%: (0.000 – 2.382)

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.

The data presented in the following reports display the status of HAI in the United States over time and currently.

The Healthcare-associated Infections in the United States, 2006-2016: A Story of Progress located here: https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html

The 2015 National and State Healthcare-associated Infection Data Report:

https://www.cdc.gov/hai/surveillance/data-reports/2015-HAI-data-report.html

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.*

Due to the imposed data entry burden, social risk factors are not collected in NHSN's MDRO surveillance module for all patients in the patient population; therefore, these variables are not available in NHSN to be used for risk adjustment modeling, and stratified data based on social disparities are not available from NHSN.

No studies provide evidence of a direct relationship between social risk and HAIs. Instead, they provide evidence that social risk factors are associated with an increased risk of chronic disease conditions, suboptimal care for those conditions, compromised functional status, exposure to nursing homes, and colonization with bacterial pathogens. While these associations may be meaningful they do not establish a direct relationship between social risk and HAIs.

Certain patient-related factors have been associated with an increased risk of MRSA.

There are studies showing patients who are found to have had direct or indirect contact with hospitals, care homes or other healthcare facilities have a higher carriage rate than those who are never exposed. Risk for infection is higher in HIV+ patients (10-17% vs. 1.2% general population) – Peters: Emerg Infect Dis. 2013 Apr19(4):623-9. doi: 10.3201/eid1904.121353.

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

(1)Klevens, RM, et al., Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. JAMA, 2007. 298(15):1763-1771.

(2) Bakullari, Anila, Mark L. Metersky, Yun Wang, Noel Eldridge, Sheila Eckenrode, Michelle M. Pandolfi, Lisa Jaser, Deron Galusha, and Ernest Moy. "Racial and Ethnic Disparities in Healthcare-Associated Infections in the United States, 2009–2011." Infection Control and Hospital Epidemiology 35, no. S3 (2014): S10-16. doi:10.1086/677827

Among patients hospitalized with acute cardiovascular disease, pneumonia, and major surgery, Asian and Hispanic patients had significantly higher rates of HAIs than white,non-Hispanic patients.

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):*

Infectious Diseases (ID)

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

Safety : Healthcare Associated Infections

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

Populations at Risk

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.cdc.gov/nhsn/pdfs/pscmanual/12pscmdro_cdadcurrent.pdf

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** NQF_1716_MRSA_attachment.docx

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.

No, this is not an instrument-based measure **Attachment:**

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.

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.

Due to changes in the NHSN protocols and population of facilities reporting data, the measure has been updated to use a new set of national baseline data from which to calculate the number of predicted events (denominator). Updating the baseline data involves creating updated risk models for each applicable healthcare setting (i.e., acute care hospitals, critical access hospitals, long term acute care hospitals, and inpatient rehabilitation facilities).

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).

Total number of observed hospital-onset unique blood source MRSA LabID events among all inpatients in the facility per NHSN protocols.

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).

1. Definition of MRSA – Includes Staphylococcus aureus cultured from any specimen that tests oxacillin-resistant, cefoxitin-resistant, or methicillin-resistant by standard susceptibility testing methods, or by a positive result from molecular testing for mecA and PBP2a; these methods may also include positive results of specimens tested by any other FDA approved PCR test for MRSA

2. Definition of MRSA isolate - Any specimen obtained for clinical decision making testing positive for MRSA. This excludes any tests related to active surveillance testing/culturing.

3. Definition of unique MRSA blood isolate - An MRSA isolate from blood in a patient that is the first MRSA isolate from any specimen for the patient in the location in that month or an MRSA isolate from blood in a patient with no prior positive blood culture for MRSA in the current inpatient location in <= 2 weeks .

4. Definition of duplicate MDRO Isolate: If monitoring MRSA , any MDRO isolate from the same patient and location after an initial isolation of the specific MDRO during a calendar month, regardless of specimen source, except unique blood source

5. Definition of MRSA Bacteremic LabID event - All non-duplicate unique blood source MRSA isolates, including specimens collected during an emergency department or other affiliated outpatient clinic visit, if collected the same day as patient admission to the facility.

6. Definition of hospital-onset LabID event – LabID event with specimen collected >3 days after admission to the hospital (i.e. on or after calendar day 4 of admission, where date of admission = day 1)

7. Definition of inpatient - A patient who is located in an inpatient location for care and treatment at the time of specimen collection. For this measure, LabID events from patients housed in a CMS-certified inpatient rehabilitation unit (IRF) or inpatient psychiatric unit (IPF) are excluded.

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

Total number of predicted hospital-onset unique blood source MRSA LabID events, calculated from a negative binomial regression model and risk adjusted for facility's number of inpatient days, inpatient community-onset MRSA prevalence rate, average length of patient stay in the hospital, medical school affiliation, facility type, number of critical care beds in the hospital, and outpatient community-onset MRSA prevalence rate from emergency departments and observation units.

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).

1. Number of inpatient days for the facility for the time period under surveillance is included in the calculation of the denominator. The number of inpatient days is obtained by summing the daily count of patients occupying beds in each applicable inpatient location in the facility over the time period under surveillance. The count of patients occupying inpatient beds is collected at the same time each day. A monthly sum of total patient days is reported to NHSN. Patient day counts from CMS-certified inpatient rehabilitation units and inpatient psychiatric units are excluded.

2. Risk factors included in the calculation of the number of predicted hospital-onset MRSA LabID events for acute care hospitals: (see attached document for further details)

- Inpatient community-onset MRSA bacteremia prevalence rate

- Average length of stay for patients in the hospital

- Medical school affiliation

- Type of hospital

-Number of ICU beds

-Community-onset prevalence rate in Emergency Departments and 24 hour observation units

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

Data from patients who are not assigned to an inpatient bed in an applicable location are excluded from the denominator counts. Denominator counts exclude data from inpatient rehabilitation units and inpatient psychiatric units with different CMS Certification Numbers (CCN) from the acute care facility.

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.)*

Definition of inpatient - A patient who is located in an inpatient location for care and treatment at the time of the daily inpatient census count.

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.)*

The measure will not be stratified, as it is an overall facility-wide summary measure. Facility characteristics will be used for risk adjustment, described above in S7.

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

Other

If other: Statistical negative binomial regression. See attachment for details.

S.12. Type of score:

Ratio

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 = Lower 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 Standardized Infection Ratio (SIR) for annual and quarterly data aggregation and analysis of MRSA bacteremia LabID events is calculated for each healthcare facility for a specified time period. The SIR is an indirect standardization method for summarizing healthcare-associated infection (HAI) experience, including MRSA bacteremia LabID events, in a single group of data or across any number of stratified groups of data. To produce the SIR:

1. Identify number of observed non-duplicate hospital-onset unique blood source MRSA LabID events for a given time period by adding the total number of observed events across the facility. Duplicate events that occurred in the same patient within a 14-day period are excluded.

2. Calculate the number of predicted hospital-onset unique blood source MRSA LabID events for the facility using the negative binomial regression model.

3. Divide the number of observed hospital-onset unique blood source MRSA LabID events (1 above) by the number of predicted hospital-onset unique blood source MRSA LabID events (2 above) to obtain the SIR.

4. Perform a mid-P Exact Test to compare the SIR obtained in 3 above to the nominal value of 1. P-value and 95% confidence intervals will be calculated, which can be used to assess statistical significance of SIR.

The Adjusted Ranking Metric (ARM) for annual data aggregation and analysis of HAI events, including MRSA bacteremia LabID events, combines the method of indirect standardization used to calculate the unadjusted SIR described above with a Bayesian random effects hierarchical model to account for the potentially low precision and/or reliability inherent in the unadjusted SIR. A Bayesian posterior distribution constructed through Monte Carlo Markov Chain sampling is used to produce the adjusted numerator. The ARM enables more meaningful statistical differentiation between hospitals by accounting for differences in patient case-mix, exposure volume (e.g. patient days, central line-days, surgical procedure volume), and unmeasured factors that are not reflected in the unadjusted SIR and that cause variation between healthcare facilities. Accounting for these sources of variability enables better measure discrimination between facilities and leads to more reliable performance rankings. To produce the ARM:

1. Identify the number of hospital-onset unique blood source MRSA LabID events for the facility

2. Obtain the adjusted number of observed hospital-onset unique blood source MRSA LabID events for the facility using a Bayesian posterior distribution constructed through Monte Carlo Markov Chain sampling which results from a Bayesian random effects model.

3. Total these numbers for an observed number of hospital-onset unique blood source MRSA LabID events

4. Obtain the predicted number of hospital-onset unique blood source MRSA LabID events (see attachment for final risk adjustment model)

5. Divide the total number of adjusted hospital-onset unique blood source MRSA LabID events (3 above) by the predicted number of hospital-onset unique blood source MRSA LabID events (4 above) to obtain the ARM.

6. Perform a Poisson test to compare the SIR obtained in 5 above to the nominal value of 1. P-value and confidence interval will be calculated, which can be used to assess significance of SIR.

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.

No sampling methodology is used in calculating the metric

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.

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

If other, please describe in S.18.

Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

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.

NHSN Laboratory-identified MDRO or CDI Event form and NHSN MDRO and CDI Prevention Process and Outcome Measures Monthly Monitoring Form

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

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

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

Facility, Other, Population : Regional and State

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

Emergency Department and Services, 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.)*

2. Validity – See attached Measure Testing Submission Form

NQF_1716_MRSA_attachment-636680193072714826.docx,NQF_MRSA_Testing_Final_based_on_NQF_feedback_resubmit_on_Aug_16.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.

No

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)

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: national healthcare safety network	☐ other:

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*).

1.3. What are the dates of the data used in testing? January 1- December 31, 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**

☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: population, region and state	☒ other: population: region and state

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*)

The standard population's hospital-onset MRSA rates that were used in the SIR baseline analysis came from all facility-wide inpatient locations (FacWideIn) reporting MRSA bacteremia LabID events to NHSN from January 1 to December 31, 2015. This represented 3,617 acute care hospitals and 14,132 facility-quarters. Hospitals were located within 55 U.S. states and territories.

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*)

Quarterly pooled NHSN hospital-level numerators ranged from 0 to 25 MRSA bacteremia LabID events (IQR 0 to 1), and the denominators ranged from 1 to 93,754 patient days (IQR 2,334 to 15,182).

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.

All data followed the same exclusion rules: quarter with zero patient days, or hospitals with missing survey variables (i.e. risk-adjustment variables) were excluded.

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.

No patient-level sociodemographic variables are used in the measure and none were available for analysis.

2a2. RELIABILITY TESTING

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

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

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

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

No additional systematic testing was conducted as the value of the measure as an indicator for differentiating good and poor performance has been substantiated by its broad use for that purpose. This measure is widely used by healthcare facilities and state health departments to inform their MRSA bacteremia surveillance and prevention efforts, by prevention collaboratives to identify intervention opportunities and measure impact of interventions, and by the Centers for Medicare and Medicaid Services for the agency's public reporting and

payment programs. In our experience, questions and concerns about the validity of MRSA bacteremia definition and criteria are infrequent and typically reflect a misunderstanding of the definition and criteria.

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*)

No additional testing was conducted.

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*)

No additional testing was performed because the measure is widely used.

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?*)

_____

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☐ **Performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

New empirical testing of the measure score has been conducted as the measure's value as an indicator for differentiating poor and good performance has been established through its wide use.

NHSN provides guidance to State Health Departments for conducting external validation of HAI data reported by facilities to NHSN within their jurisdiction. The validation process includes selection of sample of facilities and subsequently a sample of charts from the selected facilities which are reviewed by trained chart abstractors and tally against the data reported to NHSN. Case classification during the medical chart review and application of the protocol by the auditor is considered as the gold standard and compared with the facility determinations. Data accuracy measures assessed include sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity of MRSA reporting is the correct identification of MRSA positive blood specimens meeting MRSA LabID criteria as MRSA (true positive rate), whereas specificity is the correct identification of a positive MRSA blood specimens not meeting MRSA LabID criteria as "not MRSA" (true negative rate). The positive and negative predictive values (PPV and NPV respectively) are the proportions of true positive and negative MRSA's among all results that are reported by the facility during the time frame that are positive and negative, respectively.

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

		Sensitivity	**Specificity**	**Positive predictive value**	**Negative predictive value**
Tennessee	2015	80.9%	87.5%	97.5%	42.8%
Wisconsin	2009	95.2%	63.6%	93.7%	70%

New Mexico	2016	98.7%	100%	100%	98.8%
California (MRSA/VRE BSI)	2014	88%	NP*	NP*	NP*
Maine	2015	83%	NP*	74%	NP*

*NP- Not provided

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?*)

Across time, there is a range of sensitivity and specificity reported by the states performing the validation. However, the range has been smaller and higher in the most recently available time period (2014-2016) with sensitivity range of 81-99% and specificity range of 87-100%. We believe these findings are supportive of the measure's reliability.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

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*)

NHSN LabID event surveillance for MRSA bacteremia includes events identified in all inpatient units within the hospital. Those facility-quarters with zero or missing denominator data were excluded due to insufficient or missing data entry.

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*)

Exclusions:

Zero patient days for a quarter: 125 hospitals / 295 quarters

Zero admissions for a quarter: 58 hospitals / 94 quarters

Zero total annual admissions: 1 facility / 4 quarters

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*)

As no continuous variables were used in the model, we did not need to impose exclusion rules on any variable. The final variables used in the model were all categorized, and therefore there was no extreme influence by an outlier observation.

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* 2b5.**

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 6 risk factors**

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

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

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

The risk model was conducted using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. Univariate models were fist constructed to evaluate the relationship between each risk factor and the MRSA incidence rate. Details of the final multivariate risk model are below:

Table 1. MRSA Bacteremia in Acute Care Hospitals

Parameter	Parameter Estimate	Standard Error	P-value
Intercept	*-11.3759*	*0.1167*	*<0.0001*
Inpatient community-onset prevalence rate*: > 0.037 per 100 admissions	0.3650	0.0286	<0.0001
Inpatient community-onset prevalence rate*: ≤ 0.037 per 100 admissions	REFERENT	-	-
Average length of stay**: ≥ 5.1 days	0.2787	0.0343	<0.0001
Average length of stay**: 4.3-5.0 days	0.0955	0.0341	0.0050
Average length of stay**: 0-4.2 days	REFERENT	-	-
Medical school affiliation‡: Major	0.2585	0.0334	<0.0001
Medical school affiliation‡: Graduate/undergraduate	0.1166	0.0345	0.0007
Medical school affiliation‡: Non-teaching	REFERENT	-	-
Facility type: Oncology Hospital *(HOSP-ONC)*	1.1894	0.2085	<0.0001
Facility type: General Acute Care Hospital *(HOSP-GEN)*	0.4355	0.0897	<0.0001
Facility type: Other Specialty Hospital	REFERENT	-	-
Number of ICU beds‡: ≥ 45	0.5650	0.0898	<0.0001
Number of ICU beds‡: 21-44	0.4599	0.0899	<0.0001
Number of ICU beds‡: 11-20	0.3394	0.0922	0.0002
Number of ICU beds‡: 7-10	0.4720	0.0993	<0.0001
Number of ICU beds‡: 0-6	REFERENT	-	-
Outpatient community-onset prevalence rate ED/24-hour Observation Unit^: > 0.032 per 100 encounters	0.3476	0.0336	<0.0001
Outpatient community-onset prevalence rate ED/24-hour Observation Unit^: > 0 and ≤ 0.032 per 100 encounters	0.1048	0.0330	0.0015
Outpatient community-onset prevalence rate ED/24-hour Observation Unit^: 0 per 100 encounters, or no applicable locations	REFERENT	-	-

* Inpatient community-onset prevalence is calculated as the # of inpatient community-onset MRSA blood events, divided by total admissions x 100. (i.e., MRSA_admPrevBldCount /numadms * 100).
** Average length of stay is taken from the Annual Hospital Survey. It is calculated as: total # of annual patient days / total # of annual admissions.
‡ Medical school affiliation and number of ICU beds are taken from the Annual Hospital Survey.
^ Emergency department (ED)/24-hour observation unit prevalence rate combines MRSA bacteremia data from all EDs and/or 24-hour observation units into a single, de-duplicated prevalence rate. This rate is calculated as the # of unique community-onset MRSA blood events that occurred in an ED or 24-hour observation unit / total encounters * 100. (i.e., MRSA_EDOBSprevCount / numTotencounters * 100). *NOTE: If you do not have an ED or 24-hour observation location that meets the NHSN location definition and thus are not reporting MRSA bacteremia data from these locations, the number of predicted events will be risk adjusted using the referent level of this variable.*

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.

N/A

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?

Potential risk factors were selected based on availability in NHSN, literature review, and subject matter expert opinion. An expert panel from CDC DHQP was formed to identify potential risk factors in the beginning of model building process. First, all available facility-level variables from NHSN were presented to the expert panel. Facility characteristics, community-onset prevalence rates, as well as an indicator variable for cancer hospital were considered as potential risk factors.

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)**

Due to concerns about data entry burden and the paucity of evidence to support social risk factor data collection for risk adjustment purposes, social risk factors are not collected in NHSN for all patients in the patient population; therefore, these variables are not available in NHSN to be used for risk adjustment modeling.

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

Variables were eligible for entering the model at p-value=0.25 and retaining in the model at p-value=0.05 significant level. Factors were entered into a multivariate model using forward selection, based on the lowest Wald Chi-square value. Goodness of fit was assessed at each modeling step using the Akaike Information Criterion (AIC) statistics. The final model resulting from forward selection was confirmed via backwards elimination, in which each variable was sequentially removed based on the highest p-value.

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.**

N/A

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

Bootstrap sampling method was used to validate the models.

Model validation steps:

1. For each multiple logistic regression model, calculate the c-index as $C_{orginal}$.
2. Generate 100 bootstrap samples from the original dataset with the same number of records as the original sample size using sampling with replacement.
3. For each one of the new samples m=1, …,100, using the predictors of the logistic regression model from step 1 to fit the data with backward elimination approach and calculate the discrimination $as\ C_{boot}^{(m)}$. Note that the model we select from each of the m bootstrap samples could be different from the original model.
4. For each bootstrap sample, the original dataset is used for validation. For this step, the regression coefficients are fixed to their values from step 3 to determine the joint degree of over fitting from both selection and estimation. We obtain $C_{original}^{(m)}$ from this step.
5. For each one of the bootstrap samples, first we will calculate the optimism in the fit: $O^{(m)} = C_{boot}^{(m)} - C_{original}^{(m)}$. Then we obtain O by taking the average of $O^{(m)}$ from M bootstrap samples.
6. The optimism corrected performance of the original model is: $C_{adj} = C_{orginal} - O$. This value is a nearly unbiased estimate of the expected value of the optimism that would be obtained from external validation.

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

If stratified, skip to 2b3.9

Model validation results for 2015 Rebaseline: MRSA Bacteremia, Acute Care Hospitals

Parameter	Level 1	DF	Original Estimate (with 95% CI)	Pr > ChiSq	Bootstrap Estimate (with 2.5, 97.5 percentiles)	N converged
Intercept		1	-11.38 (-11.6, -11.15)	<.0001	-11.39 (-11.64, -11.14)	993
prevmed	1.High	1	0.365 (0.3088, 0.4211)	<.0001	0.3659 (0.3069, 0.4293)	993
prevmed	2.Low	0				993
LOSquart	1.Highest	1	0.2787 (0.2116, 0.3459)	<.0001	0.2788 (0.2136, 0.3426)	993
LOSquart	2.High	1	0.0955 (0.0288, 0.1623)	0.0050	0.0968 (0.0269, 0.1645)	993
LOSquart	3.Lowest	0				993
medicalschool	Major	1	0.2585 (0.193, 0.324)	<.0001	0.2572 (0.1911, 0.3264)	993
medicalschool	Minor	1	0.1166 (0.049, 0.1841)	0.0007	0.1156 (0.0475, 0.1811)	993
medicalschool	None	0				993
fac	1.ONC	1	1.1894 (0.7808, 1.598)	<.0001	1.1678 (0.5297, 1.8221)	993
fac	2.GEN	1	0.4355 (0.2597, 0.6112)	<.0001	0.4416 (0.2601, 0.6421)	993
fac	3.Special	0				993
icubedquint	1.Most	1	0.565 (0.3891, 0.7409)	<.0001	0.577 (0.3684, 0.7799)	993
icubedquint	2.Some	1	0.4599 (0.2837, 0.6361)	<.0001	0.472 (0.2691, 0.6888)	993
icubedquint	3.Middle	1	0.3394 (0.1587, 0.5201)	0.0002	0.352 (0.145, 0.5592)	993
icubedquint	4.Few	1	0.472 (0.2774, 0.6665)	<.0001	0.4814 (0.2593, 0.6954)	993
icubedquint	5.Fewest	0				993
EDMvar2	1.HiED	1	0.3476 (0.2818, 0.4135)	<.0001	0.3479 (0.2849, 0.4161)	993
EDMvar2	2.MedED	1	0.1048 (0.0401, 0.1696)	0.0015	0.1039 (0.0372, 0.1694)	993
EDMvar2	3.No_ED_Zero	0				993
Dispersion		1	0.2559 (0.2194, 0.2985)		0.253 (0.2084, 0.301)	993

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

Likelihood Ratio Test, Akaike Information Criterion and dispersion-based adjusted R-squared

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

Likelihood Ratio Test, Akaike Information Criterion and dispersion-based adjusted R-squared

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

All Likelihood Ratio Tests for the best models indicated significant improvement as well as the lowest Akaike Information Criterion values and the greatest dispersion-based adjusted R-squared.

2b3.9. Results of Risk Stratification Analysis: N/A because bootstrap method was used.

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

The p-values for all variables in the final multivariate model were statistically significant, with several variables having a p-value < 0.0001. These variable are accounting for significant differences in risk of MRSA bacteremia between healthcare facilities. With the data reported to NHSN we have made full use of the available risk factor data to produce a series of prediction models for public reporting and pay for performance.

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*)

The multivariate regression model was confirmed and validated using bootstrap validation techniques.

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 MRSA bacteremia measure data are used to calculate an observed/predicted ratio, and ratios significantly higher than 1 are indicative of a quality concern that warrants full investigation and response.

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*)

Median 0.827

National Pooled mean 0.998

N= 3,616

Significantly higher than national SIR 144 (8%)

Significantly lower than national SIR 75 (4%)

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?*)

We see variation among facilities, and we can identify the facilities for which the summary measure warrants additional investigation.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

All facilities participating in NHSN and reporting LabID events to the MDRO module follow the same protocol for reporting events using similar laboratory and admission/discharge/transfer data sources.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

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

All facilities participating in NHSN and reporting LabID events to the MDRO module follow the same protocol for reporting events using similar laboratory and admission/discharge/transfer data sources. The NHSN application provides "Alerts" to participating healthcare facilities in the event of missing data. In addition, CDC analysts conduct regular data quality checks and perform outreach to facilities regarding any missing or implausible data. Facilities that are not reporting data elements that are required by NHSN would not be eligible to receive an SIR.

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*)

Due to federal and state reporting requirements, as well as enforced business rules inside of the NHSN application, the majority of healthcare facilities are completing 100% of all required data entry, and thus minimal "missing" data exist.

Refer to 2b2.2. for exclusions related to missing data.

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

See above.

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.

Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry), Other

If other: LabID events and denominator data can be collected manually by trained hospital staff or via electronic data capture from hospital laboratory and Admission/Discharge/Transfer (ADT) systems. The SIR is automatically calculated by the NHSN web application.

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).

NHSN provides the option for facilities to collect the data electronically and download into NHSN. However, we leave the option for manual entry for facilities that are not equipped or ready to submit electronically.

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

Attachment:

3c. Data Collection Strategy

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

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

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

The NHSN Multidrug Resistant Organism and C. difficile Infection (MDRO/CDI) module has been available for facilities to use since 2009. The ability to perform facility-wide surveillance with a single denominator was introduced in 2010, reducing data collection burden on participating facilities. The ability to perform facility-

wide surveillance for MRSA LabID events from blood specimens only was also introduced in 2010 to reduce data collection burden and to offer an option to limit to invasive cases of MRSA. To further reduce case finding and data entry burden on facilities, LabID event reporting for MRSA can be performed electronically via NHSN's Clinical Document Architecture import function.

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)

Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (external benchmarking to organizations)	Public Reporting Hospital Inpatient Quality Reporting Program (HIQR) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html LTCH Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/LTCH-Quality-Reporting/index.html Hospital-Acquired Condition Reduction Program (HACRP) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HAC/Hospital-Acquired-Conditions.html Public Health/Disease Surveillance National Healthcare Safety Network http://www.cdc.gov/nhsn/ Payment Program Hospital Inpatient Quality Reporting Program (HIQR) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html LTCH Quality Reporting Program http://cms.hhs.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/LTCH-Quality-Reporting/index.html Hospital Value-Based Purchasing https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Hospital-Value-Based-Purchasing-.html Hospital-Acquired Condition Reduction Program (HACRP) https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/HAC-Reduction-Program.html Quality Improvement (Internal to the specific organization) Regulatory and Accreditation Programs https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program

	http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html

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

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

1) Name: Hospital Inpatient Quality Reporting Program (HIQR)
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: Nationwide, currently covers all acute care hospitals with ICUs (approximately 3300).*
Level of measurement and setting: Facility-Level, acute inpatient hospital
2) Name: Prospective Payment System Exempt Cancer Hospital Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for PPS-Exempt Cancer Hospital to improve health, improve care and lower cost (triple aims)of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: 11 Patient Prospective Payment Exempt Cancer Hospitals in 7 U.S. states with 19,203 average discharges each in FY 2012*.
Level of measurement and setting: Facility-Level, PPS-Exempt cancer hospital
3) Name: Inpatient Rehabilitation Facility (IRF) Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for IRFs to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: All 50 U.S. States are included, 371,288 IRF discharges in 2011*.
Level of measurement and setting: Facility-Level, acute inpatient hospital
4) Name: Long Term Care Hospital (LTCH) Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for LTCHs to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: All 442 Medicare certified long-term care hospitals are required to participate to receive 100% of reimbursement money due. In 2012, this included 202,050 patient discharges*.
Level of measurement and setting: Facility-Level, LTAC inpatient
5) Name: Hospital Value-Based Purchasing
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: 2808 entities*
Level of measurement and setting: Facility-Level, acute inpatient hospital
6) Name: Hospital-Acquired Condition Reduction Program (HACRP)
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.

Geographic area and number and percentage of accountable entities and patients included: 3,216 entities*
Level of measurement and setting: Facility-Level, acute inpatient hospital
*provided by Centers for Medicare and Medicaid Services

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?*)

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.*)

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.

Numerous training materials have been created in order to assist users with the proper understanding and interpretation of this measure. Several webinars and written training materials have been provided. Annual in-person trainings are held to discuss the SIR calculations, risk adjustment, and proper interpretation. Training materials are available online to all hospitals enrolled in NHSN, as well as external partners such as state health departments, quality improvement organizations, and healthcare corporations. NHSN users can run monthly analysis reports within NHSN to view their SIR data. On an annual basis, NHSN publishes national and state-level SIRs in the National and State HAI Progress Report.

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.

SIR results are available to NHSN users at any time, based on their current data entry. Data provided within the analysis report includes numerator, denominator, SIR, p-value, and 95% confidence interval. Educational materials are available on the NHSN website that explain each data element. NHSN provides user-support via NHSN@cdc.gov including explanations of data analysis.

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 on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided to us by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection prevention and other personnel.

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

Feedback from Hospitals and states: Based on results from a polling survey, hospitals have indicated that they are running SIR analysis reports within NHSN on a monthly basis, and that they use SIRs for prevention activities in their hospital. State health departments are using the SIR for public reporting purposes and to help target facilities for additional prevention. Feedback was received via email regarding the extent of risk adjustment and the limitation.

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

Feedback on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided to us by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection prevention and other personnel as well as consumer advocate groups.

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.

Feedback from all stakeholders is considered when developing and implementing the SIR. Different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users. Additional training formats, such as live chats and "quick learn" videos, were created in order to address different training environment that best meet the needs of our audience. We have also provided live demonstrations to users showing how to generate their SIRs in NHSN based on earlier feedback we had received.

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.

•There has been major progress since 2005 in preventing MRSA bacteremia due to declines in hospital-onset and community-onset, healthcare-associated bacteremia. However, declines are slowing in these areas.

• Using the 2015 baseline, there was a 5% decline in SIR between 2015 and 2016. There was a slow continuous decline in the unadjusted NHSN crude rate of hospital-onset MRSA bacteremia (the outcome represented by the SIR) from 2012 through 2016, ranging from 0.61 cases per 10,000 patient days to 0.55 cases per 10,000 patient days, with no increase in 2015.

https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html

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.

Laboratory and other clinical data must be reviewed to determine if the patient meets the criteria for a LabID event. It is possible that medical record reviewers will miss positive cultures or important dates that would indicate that a LabID event should be recorded. Similarly, reviewers might miss data in the medical record that would indicate a positive culture should not result in a LabID event. It is also possible that data abstractors could intentionally underreport LabID events.

Business logic is built into the NHSN application to minimize incorrect entry of LabID events. Additionally, agencies including state health departments and others have indicated interest in performing validation of LabID event surveillance as they have for other healthcare-associated infections, such as central line-associated bloodstream infections.

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

N/A

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.

No

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

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

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.)

N/A

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:** MRSA_appendix_for_NQF.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Disease Control and Prevention

Co.2 Point of Contact: Daniel, Pollock, MD, dpollock@cdc.gov, 404-639-4237-

Co.3 Measure Developer if different from Measure Steward: Centers for Disease Control and Prevention

Co.4 Point of Contact: Daniel, Pollock, MD, dpollock@cdc.gov, 404-639-4237-

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.

None

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2009

Ad.3 Month and Year of most recent revision: 07, 2018

Ad.4 What is your frequency for review/update of this measure? Annually and as needed

Ad.5 When is the next scheduled review/update for this measure? 07, 2019

Ad.6 Copyright statement: All CDC documents are public record therefore there is no copyright.

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments:

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

Brief Measure Information

NQF #: 1717

Corresponding Measures:

De.2. Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure

Co.1.1. Measure Steward: Centers for Disease Control and Prevention

De.3. Brief Description of Measure: Standardized infection ratio (SIR) and Adjusted Ranking Metric (ARM) of hospital-onset CDI Laboratory-identified events (LabID events) among all inpatients in the facility, excluding well-baby nurseries and neonatal intensive care units (NICUs).

1b.1. Developer Rationale: The SIR describes a healthcare facility´s performance compared to a national baseline. Facilities are able to see how the number of hospital-onset C. difficile LabID events they have reported compares to the number predicted, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.

S.4. Numerator Statement: Total number of observed hospital-onset incident CDI LabID events among all inpatients in the facility, excluding NICU, Special Care Nursery, babies in LDRP, well-baby nurseries, or well-baby clinics.

S.6. Denominator Statement: Total number of predicted hospital-onset CDI LabID events, calculated using the facility´s number of inpatient days, facility type, CDI event reporting from Emergency Department and 24 hour observation units, bed size, ICU bed size, affiliation with medical school, microbiological test method used to identify C. difficile, and community-onset CDI admission prevalence rate.

S.8. Denominator Exclusions: Data from patients who are not assigned to an inpatient bed are excluded from the denominator counts, including outpatient clinics, 24-hour observation units, and emergency department visits. Inpatient rehab locations and inpatient psychiatric locations that have their own Centers for Medicare and Medicaid Services (CMS) Certification Number (CCN) are excluded. Additionally, data from NICU, SCN, babies in LDRP, well-baby nurseries, or well-baby clinics are excluded from the denominator count.

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

S.20. Level of Analysis: Facility, Other, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Dec 14, 2012 **Most Recent Endorsement Date:** Dec 14, 2012

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 logic model describes how comparing the number of reported (HO) C. difficile events to the number predicted drives prevention practices (e.g., appropriate antibiotic use and isolation precautions) that lead to improved outcomes such as reduction in morbidity and mortality associated with (HO) C. difficile.
- This measure is supported by the following guidelines:
    - IDSA/SHEA Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children (2017)
    - Centers for Disease Control and Prevention´s Healthcare Infection Control Practices Advisory Committee (HICPAC) Guideline for Disinfection and Sterilization in Healthcare Facilities (2008)
    - Centers for Disease Control and Prevention´s Healthcare Infection Control Practices Advisory Committee (HICPAC) Guideline for Isolation Precautions: Preventing Transmission of Infectious Agents in Healthcare Settings (2007)
    - Guideline for Hand Hygiene in Health-Care Settings: Recommendations of the Healthcare Infection Control Practices Advisory Committee and the HICPAC/SHEA/APIC/IDSA Hand Hygiene Task (2002)
- These guidelines provide evidence for infection prevention and recommendations for practices that result in the reduction of transmission of infections within healthcare facilities, including CDI.
- The developer provided clarifying information on 12/12/18.
    - HICPAC has not published updated evidence however the IDSA/SHEA 2018 CDI Guidelines reaffirm the evidence cited in the HICPAC document. The 2015 data was more robust meaning there is more data from a greater number of facilities to validate outcomes.

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 facilities can do to achieve a change in the measure results?*
- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Outcome measure (Box 1) → Relationship between heath outcome and at least one healthcare action is demonstrated by empirical data (Box 2) → Yes → PASS

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The SIR compares a healthcare facility´s performance compared to a national baseline.
  - National CDI SIR in 2015 is 0.993 = 101,505 observed / 102,203.940 predicted
    - National % change vs. baseline in 2015 < 1%
  - National CDI SIR in in 2016 is 0.921 = 95,530 observed / 103,780.133 predicted
    - National % change vs. baseline in 2016 is 8%
  - Percent Change 2016 v. 2015 7% decrease
  - 2015
    - # facilities: 3,634
    - Median: 0.928
    - Range, at 5% and 95%: (0.000 – 1.842)
  - 2016
    - # facilities: 3,605
    - Median: 0.851
    - Range, at 5% and 95%: (0.000 – 1.729)
- Information provided indicates that the national CDI SIR as well as facility CDI SIR improved from 2015 to 2016.

Disparities

- Due to the imposed data entry burden and lack of evidence-based, analytic value for hospital-onset CDI, social risk factors are not collected in NHSN's MDRO surveillance module and are not available in NHSN for risk adjustment or stratification purposes.
- There are no studies showing a direct relationship between social factors and HAIs.

- Rates of CDI are highest for patients in healthcare facilities. Rates also increase with patient age.
    - From 1996 to 2009, C. difficile rates for hospitalized persons aged greater than or equal to 65 years increased 200%, with increases of 175% for those aged 65-74 years, 198% for those aged 75-84 years, and 201% for those aged =85 years. C. difficile rates among patients aged greater than or equal to 85 years were notably higher than those for the other age groups
- In-hospital fatality associated with C. difficile infection in the United States has decreased more than 2-fold in the last decade, despite increasing infection rates.

Questions for the Committee:

- Is the information provided enough to distinguish a gap in care? Is there additional information that would be helpful to include to identify gaps?
- Is the rationale that social risk factor data is not available and that there is no relationship between social risk and HAIs adequate?

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

RATIONALE:

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

1a. Evidence
Comments:

**The evidence is not truly new, but in aggregate supports the trageted outcome (C diff infection in hospiatlized adults). The local hygienge and antibiotic practices clearly impact the outcome and eventaul health.

**Not aware of any new information that changes the evidence base.

**Pass, an outcomes metric, clinical practice guidelines.

**Lab ID challenge there is NO gold standard for diagnosis of CDI.

**Okay.

**Adherence to preoperative measures such as avoiding shaving, sterile technique, preopdecontamination, prophylactic antibiotics, glycemic control and oxygenation can lead to decreased SSIs.

**Guidelines exist for control of infection to include isolationi precautions, hand hygiene, disinfection and sterilization techniques that when employed can decrease the indicidence of iatrogenic infection.

**No need to vote on evidence. Established.

**Pass - supported by multiple guidlines, no new literature.

**Acceptable.

**Supplied updated evidence and seems approrpiate . No need to discuss.

**NHSN facility-wide inpt hospital-onset clostridium difficile infection Population-based outcome measure. Electronic, paper and other data. The bodies of evidence reviewed in the preparation of the guidelines referenced in 1a.3 indicates recommended prevention practices can reduce incidence and transmission of CDI in healthcare settings.

1b. Performance Gap
Comments:

**Despite improvemnt, opportunity still exists and is magnified by variation that occurs in a non-systematic fashion - both can impact population health.

**Performance gap remains. No disparities in care of subgroups provided.

**Moderate, two year data showing decreasing trend, no disparities studies found linking social factors to HAIs.

**Progress for CDI is slow and soemwhat driven by lab methods and stool acceptability.

**Okay.

**SIR 0-2.5 at the facility level for hysterectomy, 0-2.17 for colon. National average 0.86 and 0.93 respectively.

**SIR improved from 2015 to 2016 by 7percent. Fatalities improved 2 fold in the last decade. Disparities show an increase in incidence with age. Current range is 0-1.729 showing variation.

**SIR 95% CI range between 0 and 1.729.

**Moderate.

**Improvement 2016-2016 and over time.

**Ample opportunity for improvement . No disparity noted but not feasible to analyze.

**The CDI measure data are used to calculate an observed/predicted ratio, and ratios significantly higher than 1 are indicative of a quality concern that warrants full investigation and response. Variation among facilities can be identified the facilities for which the summary measure warrants additional investigation. Due to the imposed data entry burden and lack of evidence-based, analytic value for hospital-onset CDI, social risk factors are not collected in NHSN's MDRO surveillance module and are not available in NHSN for risk adjustment or stratification purposes. There are no studies showing a direct relationship between social factors and HAIs.

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.

Composite measures only:

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Evaluation of Reliability and Validity:

Scientific Methods Panel Votes: Measure passes

- Reliability: H-0; M-5; L-0; I-0
- Validity: H-0; M-5; L-0; I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- The developer provided results of data element validity testing; NQF' guidance states that additional reliability testing is not needed if empirical validity testing of patient-level data is conducted and the results are adequate.
- Panel members noted that while the testing information met NQF's minimum requirements, they would have liked to see separate reliability testing of data elements.

Validity

- Validity testing was performed at the data element level.
- Data Element
    - The developers provided a summary of validation studies conducted in 5 states.
    - These studies involved taking a sample of charts from a sample of facilities in varying years; these samples were then reviewed by trained chart abstractors and compared against data reported to the National Healthcare Safety Network (NHSN).
    - Developer provides sensitivity/specificity/PPV/NPV seemingly for the Clostridium Dificile (C. Dif) variable only. Panel members also noted that testing of variables included in the risk adjustment model was not reported; no information is provided on the validity of data elements used for risk adjustment and to identify the denominator population.
    - Results:

State	Year of data validated	Records reviewed	sensitivity	specificity	PPV	NPV
Connecticut	2013	1085	93	99	99.9	75
Colorado	2015	359	95	100	100	80
Tennessee	2015	534	89.4	73.5	98	32

Utah	2016	394	92.5	100	100	42.9
New Mexico	2016	302	100	58.3	98.3	100
New York	2014	1787	89.4	100	100	42.8
Overall		**4461**	**94.9**	**93.7**	**99.4**	**58.7**

- Risk Adjustment
    - This is a risk-adjusted model with 7 risk factors: Inpatient community onset prevalence; CDI test type, medical school affiliation; number of ICU beds; facility type; facility bed size; and reporting from ED or 24-hour observation unit.
    - No social risk factors were included because these are not collected in the NHSN for all patients in the patient population.
    - The risk model was developed using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. The multivariate regression model was confirmed and validated using bootstrap validation techniques.
    - Results:
        - The p-values for all variables in the final multivariate model were statistically significant, with several variables having a p-value < 0.0001.

Standing Committee Action Item(s):

- The Standing Committee can discuss the reliability and validity testing, or agree to take the ratings of the Scientific Methods Panel.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

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

Evaluation A: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

*Note: Completed by multiple Scientific Methods Panel members and therefore multiple responses provided in checkboxes.

Measure Number: 1717

Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI) Outcome Measure

Type of measure:

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

Data Source:

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

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☒ **Population: Regional and State**
☐ **Integrated Delivery System** ☒ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

   - **PANEL MEMBER 2**: Yes, the numerator is the number of observed hospital-onset incident CI LabID events among all inpatients in the facility (with some exclusions) per NHSN protocols and the denominator is the number of predicted such events calculated from an adjusted model. This measure has been adopted nationally and applied to facilities for at least ten years.

2. **Briefly summarize any concerns about the measure specifications.**
   - **PANEL MEMBER 1:** The authors do not address situations where the infection may be present but no test ordered or errors in the lab's handling of samples.
   - **PANEL MEMBER 2:** I have no concerns about the measure specs.
   - **PANEL MEMBER 3**: none
   - **PANEL MEMBER 4**: None.
   - **PANEL MEMBER 5:** Sensitivity generally high, specificity and negative predictive value vary substantially across states. Pooled sensitivity, specificity high and 9robably adequate, but would like to see discussion of efforts to improve accuracy of reporting.

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
   - **PANEL MEMBER 2:** Previously tested, not re-tested with this submission
   - **PANEL MEMBER 3:** No testing was conducted
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
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2
   - **PANEL MEMBER 1**: The authors feel the wide use of the measure indicates that it is reliable. However, in the validity section they report on hospital audits of lab tests. This seems like a form of reliability testing.
   - **PANEL MEMBER 2**: What was provided was data regarding the sensitivity, specificity, positive predictive value, and negative predictive values across five states which were felt to be supportive of the measure's reliability.
   - **PANEL MEMBER 3:** No reliability testing submitted
   - **PANEL MEMBER 4:** "Wide use" cited as rationale for no additional testing.
   - **PANEL MEMBER 5:** The basic construction of the measure is:

     Get count of lab-based measures of CDI. Divide this by patient days to get rate.

     Estimate expected rate from negative binomial regression model.

     Construct actual to expected by dividing actual rate to estimated rate.

     Potential sources of unreliability of measure:
     a. Errors in counts of events.
     b. Variability in counts over time due to random fluctuation
     c. Errors in counts of patient day denominator.
     d. Imprecision in risk adjustment model

     Methods used:
     a. Reabstracting and assessment of sample of charts and calculation of specificity, value. Method is appropriate.
     b. Not done. Given low counts (0-25, IQR 0-1), failure to consider year to year variability due to randomness is a weakness.
     c. Not done, but this should be a de minumus source of error.
     d. Variables in regression model initially identified through expert panels based on available data taking burden into account. Continuous variables used to construct categorical variables (average LOS terciles; number of ICU beds quintiles). Other categorical variables (hospital type: cancer, general acute, other specialty; medical school affiliation: major, grad/undergrad, none). Some continuous variables (community infection rates; hospital bed size). Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on "optimism in the fit." No assessment made of stability of SIRs bootstraps or change in ranking across significantly higher (14%), significantly lower (15%)
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3
   - **PANEL MEMBER 1:** Adequate based on results described in the validity section.
   - **PANEL MEMBER 2:** Reliability testing for critical data elements was not provided with this submission. "No additional testing was conducted as the value of the measure as an indicator for

differentiating good and poor performance has been substantiated by its broad use for that purpose. The measure is widely used by healthcare facilities and state health departments".

- **PANEL MEMBER 3**: Developer did not submit reliability testing
- **PANEL MEMBER 4:** N/A.
- **PANEL MEMBER 5**: Sensitivity, specificity, positive predictive, negative predictive value vary from year to year and state to state. Probably meet minimum standards but would like to see discussion with developer of efforts to improve reporting over time.
    - Given low counts (0-98, IQR 0-10), failure to consider year to year variability due to randomness is a weakness.
    - NA
    - The developers do not report the "optimism" estimate or any other data to assess the stability of the SIRs or rankings over the bootstrapped models. The 95% CIs are wide for coefficients on some variables (e.g., 1.0681- 1.4208 for cancer hospitals, indicating a range for multiplying the rate from 3 to 4 times the reference category). There can be substantial shifts in estimated/expected depending on the risk model, with the sensitivity of the estimates or rankings to variations in estimated coefficients not presented.

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 <u>all</u> testing results):

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

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

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

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

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:** Would appreciate more information on the reliability of testing within hospitals, but the authors seem confident that state auditing is sufficient to ensure high quality testing.
- **PANEL MEMBER 2:** No additional testing performed with this submission. Reference made to widespread utilization of the measure from initial endorsement.
- **PANEL MEMBER 3:** Developer states that testing is not necessary because widely used. I do not think that is a sufficient reason to not test for reliability
- **PANEL MEMBER 4:** No additional data were provided.

- **PANEL MEMBER 5:** There is inadequate information presented on the stability of the measure to reasonable expected variation in lab reported counts of CDI or the precision and stability of the risk adjustment model.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    - **PANEL MEMBER 1:** None.
    - **PANEL MEMBER 2**: I have no concerns with exclusions which are explicitly stated.
    - **PANEL MEMBER 3**: No testing of the potential impact of missing data or the reason that data might be missing (root cause).
    - **PANEL MEMBER 5**: None.

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

    Submission document: Testing attachment, section 2b4.

    - **PANEL MEMBER 1:** The authors point out that the measure varies between facilities cross-sectionally and over time. They feel the differences are meaningful and suggest the wide use of the measure is an indication of its value to providers and policy makers. More data on the within versus between facility variation in the measure would have been helpful.
    - **PANEL MEMBER 2:** Citation is given that the median SIR score is 0.928 and the national pooled mean in 0.993 amongst 3,634 facilities. There are 14% higher than the national SIR and 15% lower than the national SIR.
    - **PANEL MEMBER 3:** Level of significance not reported in 2b4.2 – not clear what 'significantly higher/lower' really means without that statistic.
    - **PANEL MEMBER 4**: None.
    - **PANEL MEMBER 5:** See discussion re reliability in items 7a,d, 11.

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

    - **PANEL MEMBER 3:** No concerns as the data sources are the same
    - **PANEL MEMBER 4:** N/A.
    - **PANEL MEMBER 5:** NA

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

    Submission document: Testing attachment, section 2b6.

    - **PANEL MEMBER 2:** No concerns. Pre-emptive alerts are given for missing data, and if not corrected, the facility does not receive a SIR
    - **PANEL MEMBER 4:** None.
    - **PANEL MEMBER 5:** NA

16. **Risk Adjustment**

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

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

    ☒ Yes ☐ No ☒ Not applicable

    16c. **Social risk adjustment:**

    16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☒ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☒ No

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

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

- **PANEL MEMBER 1:** The model has good face validity, is parsimonious and good statistical properties.
- **PANEL MEMBER 2:** The risk adjustment model was conducted using a negative binomial regression. Univariate models were first constructed to assess relationships between the risk factor and the CDI incidence rate, then applied to a multivariate model. Selection criteria were eligibility for inclusion at a p value of 0.25 and retention at a p value of 0.05. In the multivariate model, forward selection was utilized based on the lowest Wald Chi-square value. Goodness of fit was applied at each modeling step using the AIC statistics. The final model was then confirmed using backwards elimination, starting with the highest p value. Model validation was tested by a bootstrap sampling method and the results are provided. The values for all variables in the final multivariate model were statistically significant, with several less than 0.0001.
    - Potential risk factors were selected based on availability in the source database, NHSN, literature review, and subject matter expert opinion. Social risk factors were not specifically included due to data entry burden and a cited lack of evidence that supports the hypothesis that data collection of such would justify inclusion.
- **PANEL MEMBER 3**: Risk adjustment approach is appropriate,
- **PANEL MEMBER 5**: Variables in regression model initially identified through expert panels based on available data taking burden into account. Continuous variables used to construct categorical variables (community infection rate dichotomized; average LOS terciles; number of ICU beds quintiles). Other categorical variables (hospital type: cancer, general acute, other specialty; medical school affiliation: major, grad/undergrad, none). Model testing for selection of variables is described, but variables considered and not in model or alternative specification of continuous variables not presented. Statistical tests used to assess final risk adjustment model described but values not presented. Stability of model estimates assessed by doing 100 bootstrapped regressions, computing coefficients and estimating c statistics from original and bootstrapped models. Assess performance based on “optimism in the fit.” No assessment made of stability of SIRs across bootstraps or change in ranking across significantly higher, significantly lower, although only 8% in higher and 4% in lower category.
    - The developers do not report the “optimism” estimate or any other data to assess the stability of the SIRs or rankings over the bootstrapped models. The 2.5-97.5% CIs are wide for coefficients on some variables. There can be substantial shifts in estimated/expected depending on the risk model, with the sensitivity of the estimates or rankings to variations in estimated coefficients not presented.
    - I’m willing to accept the variables included in the model as a potentially reasonable basis for differentiating expected performance across hospitals. The community acquired rates seem like a useful adjuster. There is, however, inadequate information presented on the precision and stability of the risk adjustment model.
    - There is not sufficient evidence presented for the measure to be used for interhospital comparisons or ranking.

- Additional information provided by the developer on 12/12/2018 in response to clarification around rationale for not using patient-level characteristics for adjustment
    - The ability to perform facility-wide surveillance with a single denominator was introduced in 2010, reducing data collection burden on participating facilities. In order to provide patient-level risk adjustment, patient level information, such as age, would need to be provided not only for the event data, i.e. CDI LabID Event, as it is now but also for the accompanying denominator data. This would mean that age would need to be associated to each patient day, patient admission and patient encounters for involved patient locations. At this time denominator data is not collected with patient level information as it would be extremely burdensome on the part of facilities. For that reason, the measure is not adjusted according to patient-level data.

VALIDITY: TESTING

17. **Validity testing level: ☐ Measure score ☒ Data element ☐ Both**
18. **Method of establishing validity of the measure score:**
    - ☐ **Face validity**
    - ☒ **Empirical validity testing of the measure score**
    - ☒ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    - **PANEL MEMBER 1**: Reviewed chart review data provided by 6 states.
    - **PANEL MEMBER 2:** What was provided was data regarding the sensitivity, specificity, positive predictive value, and negative predictive values across six states which were felt to be supportive of the measure's reliability.
    - **PANEL MEMBER 3**: Validity testing in 5 states from 2009 to 2016
    - **PANEL MEMBER 4:** Adequate sensitivity/specificity, PPV/NPV data provided, although variable by state.
    - **PANEL MEMBER 5**: Chart review of numerator.
20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    - **PANEL MEMBER 1:** Very high sensitivity and specificity
    - **PANEL MEMBER 2:** The process and data provided in Q16 above is used to establish validity of the model to derive the CDI score
    - **PANEL MEMBER 3:** Difficult to assess – there is not sample size listed for the states nor is there any statistical testing provided.
    - **PANEL MEMBER 4**: See 19.
    - **PANEL MEMBER 5**: Estimates of sensitivity, specificity, positive and negative predicted value from chart review of sample.
21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

☒ **Yes**

☐ **No**

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

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

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

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

☐ **Low** (NOTE: Should rate LOW if you believe that there 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.)

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

- **PANEL MEMBER 1**: More analysis at the population or group level would be helpful.
- **PANEL MEMBER 2:** A formal testing of the score is not provided. What is provided in the rationale for model development and data showing national and regional differences.
- **PANEL MEMBER 4**: Analyses provided indicate adequate validity for level being tested.
- **PANEL MEMBER 5:** Ideally, sensitivity, specificity, positive and negative predictive values would be higher, particularly specificity and negative predictive value, but within range of acceptability.

ADDITIONAL RECOMMENDATIONS

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

- **PANEL MEMBER 2:** This was a very difficult measure to assess. It has strong national usage and application but the document provided does not provide either interval information regarding testing or the details needed to make thorough assessment. Much is deferred to a prior endorsement. The steering committee assessment of 2012 is provided and is noted to be high on importance, moderate+ on reliability and validity, high on usability and high on feasibility. The measure does provide extensive detail on the model's development for risk stratification.
- **PANEL MEMBER 5**: The standing committee should discuss the basis for expert panel assessment of variables to be included in risk adjustment and extent to which variations in results across potential risk adjustment models should be presented.
    - Question should also be discussed about what types of analysis should be presented to allow assessment of year to year stability of the measure (reliability) and consistency of scoring and ranking when comparison is made across hospitals.

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

2a1. Reliability – Specifications
Comments:

**No concerns - the underpinning seems approriate and accurate.

**No concerns.

**Specifications provided.

**Data elements and risk adjustment clearly defined but may not capture all elements.

**I think there is room for more clarity as to how to diagnose and when.

**Exclusions are not clear. What are off plan colon and abdominal hysterectomies.

**Specifications are clear with positive lab ID required. Does not address if no lab was ordered.

**Agree with prelim moderate rating.

**Moderate.

**Acceptable.

**No concerns.

**The developer provided results of data element validity testing; NQF' guidance states that additional reliability testing is not needed if empirical validity testing of patient-level data is conducted and the results are adequate. Panel members noted that while the testing information met NQF's minimum requirements, they would have liked to see separate reliability testing of data elements.

2a2. Reliability – Testing
Comments:

**No.

**None.

**Moderate, data element validation conducted.

**No.

**Yes especially in age less than 2 years.

**Colon mean 50% with 1/3 below 40%. Hysterectomy mean 52.9% with majority exceeding 40%.

**Data element testing done for sensitivity, specificity, positive predictive value and negative predictive value across 5 states. No evaluation of year over year variation due to randomness. Given small numbers this could impact outcomes. Tests on risk adjustment model described but not presented.

**No.

**None.
**No.
**None.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:

**No.

**No concerns.

**Moderate, results provided.

**No.

**Depending upon screening test.

**Unclear which data elements were tested. low sensitivity. No testing of risk adjustment.

**Year over year variation due to randomness because of the small number could impact the outcomes.

**No.

**None.

**No.

**None.

**Developer provides sensitivity/specificity/PPV/NPV seemingly for the Clostridium Dificile (C. Dif) variable only. Panel members also noted that testing of variables included in the risk adjustment model was not reported; no information is provided on the validity of data elements used for risk adjustment and to identify the denominator population.

**Given the mesaure maturity, these data are clear; the threats do not appear to grow or engulfing of real opportunity

**No concerns.

** No concerns.

** Problem is what test is used and if facility has method to make sure appropirate stools are submitted.

**No.

**Many facilities did not meet minimum precision criteria. Because of the low frequency this may not reflect true comparison.

** Infections that were not cultured will not show so skewing the results.

**No.

**None

**No concerns.

**No.

**No threats.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**I do not see transplant or chemotherapy factors accounted for - sites that had larger subpopulations of these may have a different performance despite good practice. The medical school weighting does not fully address this concern.

**No concerns.

**Risk adjusted model with 7 risk factors.

**Same as above.

**Is there a way to capture management or patient co-morbidities better.

**Concerns about exclusions which appear vague - data quality, outlier, errors. Again what does off plan mean?

**Need better understanding of variables used for risk adjusting and validity of these.

**None – moderate.

**Appreciate the limitations of NHSN re: social determinants.

**No fatal threats.

**The model has good face validity, is parsimonious and good statistical properties.

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.

- All data elements are in defined fields in a combination of electronic sources.
- NHSN provides the option for facilities to collect the data electronically and download into NHSN. They leave the option for manual entry for facilities that are not equipped or ready to submit electronically.

Questions for the Committee:

- Are there any difficulties the Committee is aware of regarding the feasibility of the measure?

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**No concerns.

**No concerns about feasibility.

**High, data collected through medical records or NHSN electronic form.

**No concerns.

**Again, better definition of diagnosis.

**Data elements generated during the provision of care some in electronic sources. NHSN tool available.

**Agree with high feasibility prelim rating.

**High.

**No concerns.

**No concerns.

**All data elements are in defined fields in a combination of electronic sources. NHSN provides the option for facilities to collect the data electronically and download into NHSN. They leave the option for manual entry for facilities that are not equipped or ready to submit electronically.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

- The measure is used in numerous public reporting and payment programs: Hospital Inpatient Quality Reporting Program (HIQR), Prospective Payment System Exempt Cancer Hospital Quality Reporting Program, Inpatient Rehabilitation Facility (IRF) Quality Reporting Program, Long Term Care Hospital (LTCH) Quality Reporting Program, Hospital Value-Based Purchasing, and Hospital-Acquired Condition Reduction Program (HACRP).
- The measure is also used for Public Health/Disease Surveillance.

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

- Feedback to the developer is provided by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection prevention and other personnel.
    - Based on results from a polling survey, hospitals have indicated that they are running SIR analysis reports within NHSN on a monthly basis, and that they use SIRs for prevention activities in their hospital.
    - State health departments are using the SIR for public reporting purposes and to help target facilities for additional prevention.
- Feedback was received regarding the extent of risk adjustment and the limitations.
    - Marra et al. article contends that the NHSN CDI LabID event healthcare quality performance measure does not adequately risk adjust for CDI test method.

        - Per developer, the authors are not clear about the appropriate litmus test for judging the adequacy of risk adjustment and their critique of the NHSN risk adjustment is based on faulty premises and a flawed analytic strategy.
- In response to feedback:
    - NHSN updated protocol to use final lab test result when performing a multistep testing methodology for CD identification. The final result of the last test finding which is placed onto the patient medical record will determine if the CDI laboratory assay definition is met, enabling use of test result which better reflects clinical determination.
    - Different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users.
    - Additional training formats and demonstrations were created to assist with understanding, interpreting, and implementing this measure.
- This measure meets all three criteria for feedback, although additional details regarding the feedback received (number of measured entities given measure results/assistance with implementation or interpretation, number of entities that provided feedback) would be helpful.

Additional Feedback: N/A

***Questions for the Committee**:*

- Are there concerns about the feedback provided in the Marra et al. article?
- Has the measure been appropriately vetted in real-world settings by those being measured?

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

RATIONALE:

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

- By the end of 2014 there had been an 8% decline in the SIR from baseline.
- Between 2015 and 2016 there was another 8% decline using the 2011 baseline and 7% decline using the 2015 baseline.
- The most recent pace of progress needs to remain steady or increase to meet national prevention goals for hospital-onset CDI in 2020.
- Crude rates of healthcare-associated CDI are decreasing, which largely reflects declines in nursing home-onset infections, along with some declines in hospital-onset CDI.

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

N/A

Potential harms

- It is possible that medical record reviewers will miss positive cultures or important dates that would indicate that a LabID event should be recorded.

- Reviewers might miss data in the medical record that would indicate a positive culture should not result in a LabID event.
- It is possible that data abstractors could intentionally underreport LabID events.
- Business logic is built into the NHSN application to minimize incorrect entry of LabID events.
- Agencies have indicated interest in performing validation of LabID event surveillance

Additional Feedback: N/A

***Questions for the Committee**:*

- During the previous measure evaluation, the Committee noted that the use of antibiotics to treat CDI could be susceptible to overuse and misuse. Is there any evidence that this is happening and is this still a concern?
- Is there concern of any other unintended impacts on patients?
- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

RATIONALE:

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

4a1. Use - Accountability and Transparency
Comments:
**No concerns - wide use and sharing.
**Already in use.
**High, publically reported in payment programs.
**Publically reported but performance challenged by above discussions.
**Need to link to management.
**Currently used in hospital inpatient quality reporting program, cancer hospital reporting program, hospital value based purchasing and HAC.
**This is currently being used for quality reporting and payment programs suchs as HIQR, IRF, LTCH, hospital based purchasing and HACRP. Feedback shows that hospitals use this for.
**Agree with prelim rating.
**Pass.
**Used in payment programs.
**Pass.
**Publicly reported and current use in accountability program. The measure is used in numerous public reporting and payment programs: Hospital Inpatient Quality Reporting Program (HIQR), Prospective Payment System Exempt Cancer Hospital Quality Reporting Program, Inpatient Rehabilitation Facility (IRF) Quality Reporting Program, Long Term Care Hospital (LTCH) Quality Reporting Program, Hospital Value-Based Purchasing, and Hospital-Acquired Condition Reduction Program (HACRP). The measure is also used for Public Health/Disease Surveillance.

4b1. Usability – Improvement
Comments:
** Evidence that these existing data altered practice and outcomes support the measure.
**Benefits seem to outweigh any unintended consequences.
**High, some risk of underreporting and possibly some missed data.
**Unintended consequences facilities may game system by changing methods although there is a risk adjustment.

**No.

**Feedback regarding risk adjustment and limitations. Risk of under reporting or miscalculating SIRs.

**Performance results can direct attention to areas where interventions might limit transmission of Cdif. There may be underreporting and perhaps under treatment associated with this.

**Minimal apparent unintended consequences.

**Pass.

**No unintended consequences.

**No concerns.

**By the end of 2014 there had been an 8% decline in the SIR from baseline. Between 2015 and 2016 there was another 8% decline using the 2011 baseline and 7% decline using the 2015 baseline. The most recent pace of progress needs to remain steady or increase to meet national prevention goals for hospital-onset CDI in 2020.

Criterion 5: Related and Competing Measures

Related or competing measures

1716: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure

Harmonization

These measures appear to be harmonized to the extent possible.

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

5. Related and Competing

Comments

**No.

**No.

**None.

**No.

**Antibiotic stewardship?

**Harmonized not competing.

**No.

**Related to 1716 – MRSA.

**Harmonized.

**No concerns.

**Related but again different organism (MRSA vs C-Diff).

**1716: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Methicillin-resistant Staphylococcus aureus (MRSA) Bacteremia Outcome Measure.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

Public Comment

** The Federation of American Hospitals (FAH) appreciates the opportunity to comment on NQF #1717: National Healthcare Safety Network (NHSN) Facility-wide Inpatient Hospital-onset Clostridium difficile Infection (CDI). FAH requeststhat the Patient Safety Standing Committee consider whether sufficient information has been provided regarding the data element validity testing under Criterion 2b. Validity. The measure developer notes that the validation was completed on a sample of hospitals and patient charts in each state but we were unable to determine whether the sampling was sufficient and question whether the information aggregated at the state level rather than for each facility and at the measure score and not for each individual data element demonstrates valid data capture and reporting at the facility level. We believe that additional information to demonstrate the validity of each data element by facility is needed to meet the validity criterion.

Support/Non-Support

- No NQF Members have submitted support/non-support choices as of this date.

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

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

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

NQF_CDI_Evidence_Final_July_27-636683050992403800.docx,1717_Evidence_MSF5.0_Data-635278463294549427-636687181599183379.doc

1a.1 <u>For Maintenance of Endorsement:</u> 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*)**:** 1717

Measure Title: National Healthcare Safety Network (NHSN) Facility-wide Inpatient hospital-onset Clostridium difficile Infection (CDI)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 7/27/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: 3 Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: 5 a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured structure leads to a desired health outcome.
- Efficiency: 6 evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☒ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

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

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

☐ Process: Click here to name what is being measured

☐ Appropriate use measure: Click here to name what is being measured

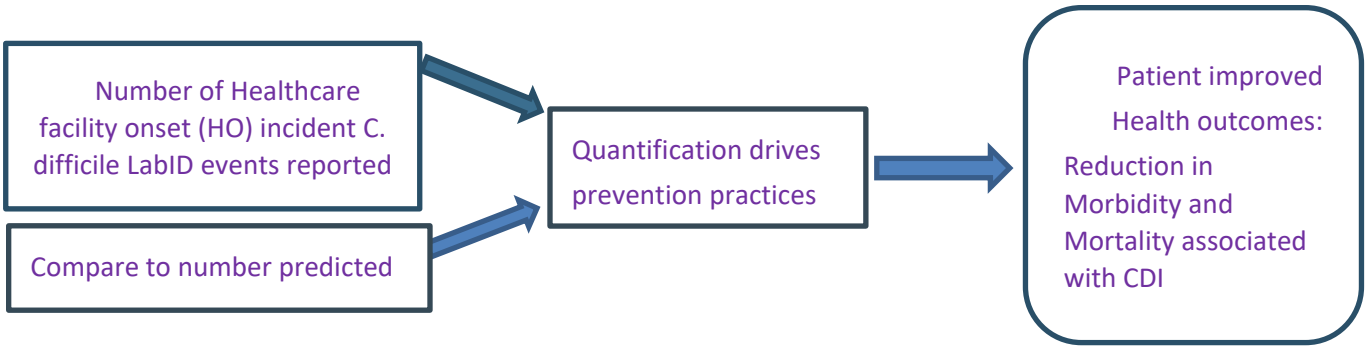
☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram

should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The SIR describes a healthcare facility´s performance compared to a national baseline. Facilities are able to see how the number of hospital-onset (HO) *C. difficile* LabID events they have reported compares to the number predicted, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.



*Measurement leads to appropriate antibiotic use, isolation precautions, HO incidence rate decline.

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

The bodies of evidence reviewed in the preparation of the guidelines referenced in 1a.3 indicates recommended prevention practices can reduce incidence and transmission of CDI in healthcare settings.

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

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

For the 2017 clinical guideline for management of CDI, an expert review panel from the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology (SHEA) of America graded existing evidence. The IDSA/SHEA guideline for management of CDI uses a standard process that includes a weighing of quality of evidence for practices that lead to successful management of CDI in the inpatient setting.

A panel of experts was convened by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA) to update the 2010 clinical practice guideline on *Clostridium difficile* infection (CDI) in adults. The update, which has incorporated recommendations for children (following the adult recommendations for epidemiology, diagnosis, and treatment), includes significant changes in the management of this infection and reflects the evolving controversy over best methods for diagnosis. *Clostridium difficile* remains the most important cause of healthcare-associated diarrhea and has become the most commonly identified cause of healthcare-associated infection in adults in the United States. Moreover, *C. difficile* has established itself as an important community pathogen. Although the prevalence of the epidemic and virulent ribotype 027 strain has declined markedly along with overall CDI rates in parts of Europe, it remains one of the most commonly identified strains in the United States where it causes a sizable minority of CDIs, especially healthcare-associated CDIs. This guideline updates recommendations regarding epidemiology, diagnosis, treatment, infection prevention, and environmental management.

The Centers for Disease Control and Prevention´s Healthcare Infection Control Practices Advisory Committee (HICPAC) graded evidence for the disinfection/sterilization, isolation precautions, and hand hygiene guidelines.

HICPAC is a federal advisory committee made up of 14 external infection control experts who provide advice and guidance to the Centers for Disease Control and Prevention (CDC) and the Secretary of the Department of Health and Human Services (HHS) regarding the practice of health care infection control, strategies for

surveillance and prevention and control of health care associated infections in United States health care facilities.

The HICPAC guidelines for sterilization and disinfection, isolation precautions, and hand hygiene provide recommendations for practices that result in the reduction of transmission of infections within healthcare facilities, including CDI. As is standard with all HICPAC guidelines, recommendations were categorized on the basis of existing scientific data, theoretical rationale, applicability, and economic impact.

Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA)

L. Clifford McDonald,1 Dale N. Gerding,2 Stuart Johnson,2,3 Johan S. Bakken,4 Karen C. Carroll,5 Susan E. Coffin,6 Erik R. Dubberke,7Kevin W. Garey,8 Carolyn V. Gould,1 Ciaran Kelly,9 Vivian Loo,10 Julia Shaklee Sammons,6 Thomas J. Sandora,11 and Mark H. Wilcox12

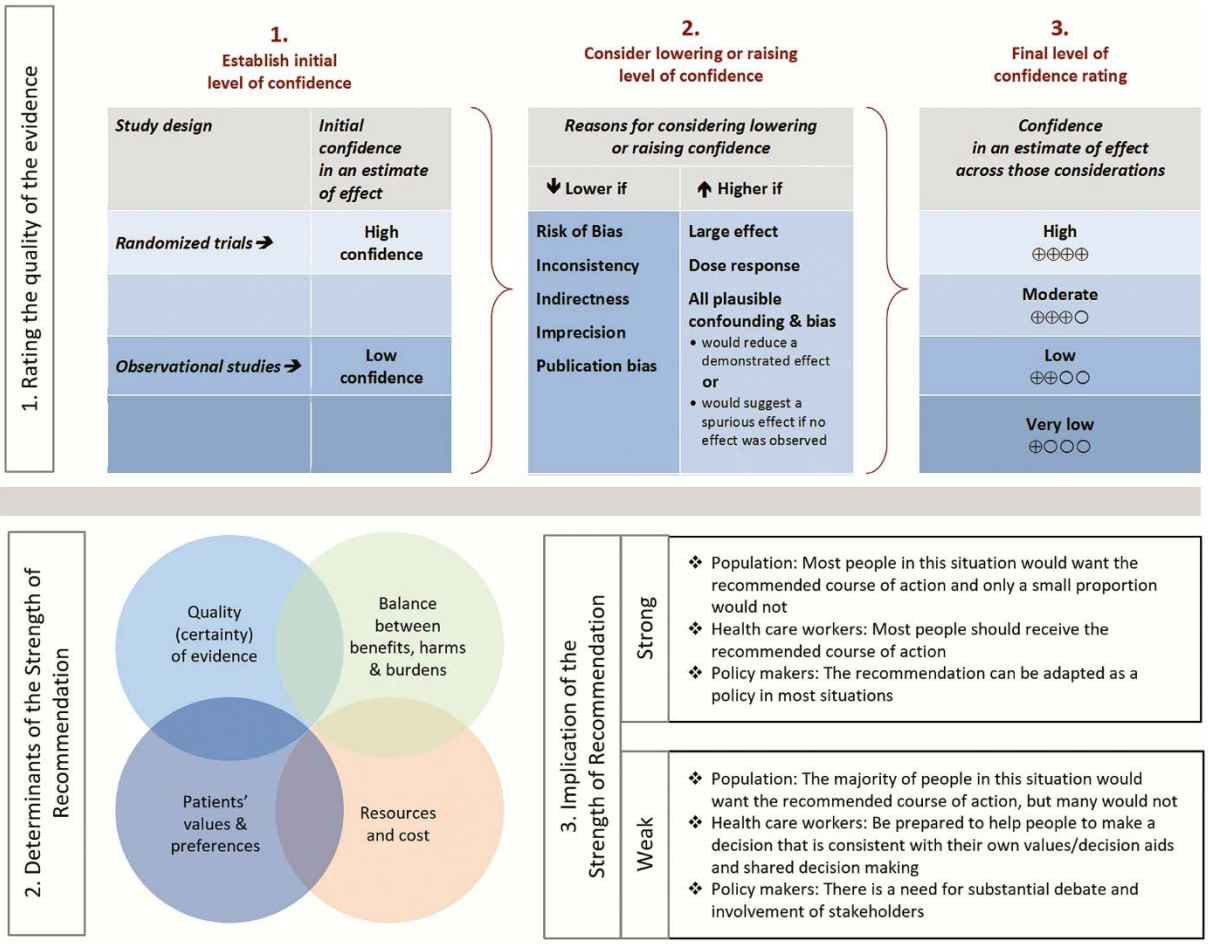
***Clinical Infectious Diseases*, Volume 66, Issue 7, 19 March 2018, Pages e1–e48,**

Published:15 February 2018

https://doi.org/10.1093/cid/cix1085

One specific overall guideline recommendation is not provided. Each individual recommendation in a guideline is given a grade as described below:

The panel followed a process used in the development of other Infectious Diseases Society of America (IDSA) guidelines, which included a systematic weighting of the strength of recommendation and quality of evidence using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) system



"A targeted systematic review of the literature was conducted in MEDLINE, EMBASE, CINAHL, and the Cochrane Library from 1998 through April 2014. A modified Grading of Recommendations, Assessment,

Development, and Evaluation (GRADE) approach was used to assess the quality of evidence and the strength of the resulting recommendation and to provide explicit links between them. Of 5759 titles and abstracts screened, 896 underwent full-text review by 2 independent reviewers. After exclusions, 170 studies were extracted into evidence tables, appraised, and synthesized".

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000094

The 2018 updated IDSA/SHEA/ practice guideline for the management of CDI included results from over 300 studies.

Rutala WA, Weber DJ, and the Healthcare Infection Control Practice Advisory Committee. **Guideline for Disinfection and Sterilization** in Healthcare Facilities, 2008. Available at https://www.cdc.gov/infectioncontrol/pdf/guidelines/disinfection-guidelines.pdf

Siegel JD, Rhinehart E, Jackson M, Chiarello L, and the Healthcare Infection Control Practices Advisory Committee. 2007 **Guideline for Isolation Precautions**: Preventing Transmission of Infectious Agents in Healthcare Settings. Available at https://www.cdc.gov/infectioncontrol/pdf/guidelines/isolation-guidelines.pdf

Boyce JM, Pittet D, et al. **Guideline for Hand Hygiene** in Health-Care Settings: Recommendations of the Healthcare Infection Control Practices Advisory Committee and the HICPAC/SHEA/APIC/IDSA Hand Hygiene Task Force. MMWR, 2002. 51(RR-16).

https://www.cdc.gov/mmwr/PDF/rr/rr5116.pdf

One specific overall guideline recommendation is not provided in any of the HICPAC guidelines. Each individual recommendation in a guideline is given a grade as described below.

Example from Siegel

Monitor the incidence of epidemiologically-important organisms and targeted HAIs that have substantial impact on outcome and for which effective preventive interventions are available; use information collected through surveillance of high-risk populations, procedures, devices and highly transmissible infectious agents to detect transmission of infectious agents in the healthcare facility (Grade 1A)

There is no overall grade assigned to the guideline, which contains many recommendations. Individual recommendations are graded

The CDC/HICPAC grading system used in the creation of the sterilization/disinfection, isolation precautions, and hand hygiene guidelines is as follows:

Category IA - strongly recommended for implementation and strongly supported by well-designed experimental, clinical, or epidemiologic studies.

Category IB - strongly recommended for implementation and supported by some experimental, clinical, or epidemiologic studies and a strong theoretical rationale.

Category IC - required for implementation, as mandated by federal and/or state regulation or standard.

Category II - suggested for implementation and supported by suggestive clinical or epidemiologic studies for a theoretical rationale.

No recommendation - unresolved issue. Practices for which insufficient evidence or no consensus regarding efficacy exists.

As above in evidence grading system

The 2008 HICPAC guideline for sterilization and disinfection in healthcare facilities included results from over 1,000 studies.

The 2007 HICPAC guideline for isolation precautions in healthcare facilities included results from over 1,100 studies.

The 2002 HICPAC guideline for hand hygiene in healthcare settings included results from over 400 studies.

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

N/A

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

1a.4.3. Provide the citation(s) for the 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 SIR describes a healthcare facility´s performance compared to a national baseline. Facilities are able to see how the number of hospital-onset C. difficile LabID events they have reported compares to the number predicted, given national data. The measure can then be used to drive prevention practices that will lead to improved outcomes, including the reduction of patient morbidity and mortality.

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.*

National CDI SIR in 2015 is 0.993 = 101,505 observed / 102,203.940 predicted

National % change vs. baseline in 2015 is < 1%

National CDI SIR in in 2016 is 0.921 = 95,530 observed / 103,780.133 predicted

National % change vs. baseline in 2016 is 8%

Percent Change 2016 v. 2015 7% decrease

2015-

facilities: 3,634

Median: 0.928

Range, at 5% and 95%: (0.000 – 1.842)

2016-

facilities: 3,605

Median: 0.851

Range, at 5% and 95%: (0.000 – 1.729)

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.

The data presented in the following reports display the status of HAI in the United States over time and currently.

The Healthcare-associated Infections in the United States, 2006-2016: A Story of Progress located here: https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html

The 2015 National and State Healthcare-associated Infection Data Report: https://www.cdc.gov/hai/surveillance/data-reports/2015-HAI-data-report.html

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.*

Due to the imposed data entry burden and lack of evidence-based, analytic value for hospital-onset CDI, social risk factors are not collected in NHSN's MDRO surveillance module for all patients in the patient population and these variables are not available in NHSN for risk adjustment or stratification purposes.

No studies provide evidence of a direct relationship between social risk and HAIs. Instead, they provide evidence that social risk factors are associated with an increased risk of chronic disease conditions, suboptimal care for those conditions, compromised functional status, exposure to nursing homes, and colonization with bacterial pathogens. While these associations may be meaningful they do not establish a direct relationship between social risk and HAIs.

"Rates of CDI are highest for patients in healthcare facilities. Rates also increase with patient age."

QuickStats: Rates of Clostridium difficile Infection Among Hospitalized Patients Aged =65 Years, by Age Group - National Hospital Discharge Survey, United States, 1996-2009. MMWR, 2011. 60(34):1171.

Clostridium difficile infections can lead to diarrhea, sepsis, and even death. The majority of infections with C. difficile occur among persons aged =65 years and among patients in health-care facilities, such as hospitals and nursing homes. From 1996 to 2009, C. difficile rates for hospitalized persons aged greater than or equal to 65 years increased 200%, with increases of 175% for those aged 65-74 years, 198% for those aged 75-84 years, and 201% for those aged =85 years. C. difficile rates among patients aged greater than or equal to 85 years were notably higher than those for the other age groups

Clostridium difficile infection discharges increased from 19.9 per 100,000 persons in 2004 to 33.8 per 100,000 persons in 2014. Clostridium difficile-associated fatality decreased from 3.6% in 2004 to 1.6% in 2014 (P < .001). Among patients aged 45-64 years, fatality decreased from 1.2% in 2004 to 0.7% in 2014 (P < .001). Among patients aged 65-84 years, fatality decreased from 4.3% in 2004 to 2.0% in 2014 (P < .001). Among patients aged =85 years, fatality decreased from 6.9% in 2004 to 3.6% in 2014 (P < .001). The mean length of hospital stay decreased from 6.9 days in 2004 to 5.8 days in 2014 (P < .001). The mean hospital charges increased from 2004 ($24,535) to 2014 ($35,898) (P < .001).

CONCLUSION:

In-hospital fatality associated with C. difficile infection in the United States has decreased more than 2-fold in the last decade, despite increasing infection rates.

https://www.ncbi.nlm.nih.gov/pubmed/28801226

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

N/A

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):*

Infectious Diseases (ID)

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

Safety : Healthcare Associated Infections

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

Populations at Risk

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.cdc.gov/nhsn/pdfs/pscmanual/12pscmdro_cdadcurrent.pdf

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** NQF_CDI_ACH_attachment_2018_Final-636692505821528619.docx

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.

No, this is not an instrument-based measure **Attachment:**

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.

Not an instrument-based measure

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.

Due to changes in the NHSN protocols and population of facilities reporting data, the measure has been updated to use a new set of national baseline data from which to calculate the number of predicted events (denominator). Updating the baseline data involves creating updated risk models for each applicable healthcare setting (i.e., acute care hospitals, critical access hospitals, long term acute care hospitals, and inpatient rehabilitation facilities).

To reduce subjectivity, NHSN protocols were updated in January 2018 to change the definition of CDI positive laboratory assay (measure numerator) to include only those specimens that tested positive from the final C. difficile test. This change applies only to those facilities performing a multi-step CDI testing algorithm for the detection of C. difficile using the same unformed stool specimen.

A reporting note was added to the 2018 Protocol to clarify difference in testing algorithms available for use at the facility level. The reporting note indicates the final test performed is to be used in meeting definition.

The 2018 MDRO/CDI Protocol states that the results of the final test that are placed in the patient's medical record should be used to determine whether or not the event meets the CDI LabID Event definition.

This means that facilities using a multi-step testing algorithm for C. difficile will be entering LabID events in NHSN based on the results of the final test in the algorithm; therefore, the standardized infection ratios (SIRs) for these facilities should be risk adjusted based on the final test in the testing algorithm. NHSN already does this for almost all multi-step algorithms listed on the FacWideIN denominator form.

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).

Total number of observed hospital-onset incident CDI LabID events among all inpatients in the facility, excluding NICU, Special Care Nursery, babies in LDRP, well-baby nurseries, or well-baby clinics.

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).

1. Definition of CDI-positive laboratory assay - A positive laboratory test result for C. difficile toxin A and/or B or a toxin-producing C. difficile organism detected by culture or other laboratory means performed on an unformed stool sample. When using a multi-testing methodology for CD identification, the final result of the last test finding which is placed onto the patient medical record will determine if the CDI laboratory assay definition is met.

2. Definition of duplicate CDI-positive test - Any C. difficile toxin-positive laboratory result from the same patient and location, following a previous C. difficile toxin-positive laboratory result within the last 14 days.

3. Definition of CDI LabID event - All non-duplicate C. difficile toxin-positive laboratory results, including specimens collected in an emergency department or 24-hour observation location.

4. Definition of hospital-onset LabID event – LabID event with specimen collected >3 days after admission to the hospital (i.e. on or after calendar day 4 of admission, where date of admission = day 1)

5. Definition of inpatient - A patient who is located in an inpatient location for care and treatment at the time of specimen collection.

6. Definition of incident CDI LabID Event - Any CDI LabID Event from a specimen obtained > 56 days after the most recent CDI LabID Event (or with no previous CDI LabID Event documented) for that patient. Note: the date of first specimen collection is considered day 1.

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

Total number of predicted hospital-onset CDI LabID events, calculated using the facility´s number of inpatient days, facility type, CDI event reporting from Emergency Department and 24 hour observation units, bed size,

ICU bed size, affiliation with medical school, microbiological test method used to identify C. difficile, and community-onset CDI admission prevalence rate.

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).

1. Number of inpatient days for the facility for the time period under surveillance. The number of inpatient days is obtained by summing the daily count of patients occupying beds in each inpatient location in the facility over the time period under surveillance. The count of patients occupying inpatient beds is collected at the same time each day.

2. Facility–specific information, including facility type, bed size, number of ICU beds, and affiliation with a medical school (see 3 below).

3. Medical school affiliation categories:

a. Major – facility has a program for medical students and post-graduate medical training

b. Graduate – facility has a program for post-graduate medical training (i.e., residency and/or fellowships)

c. Undergraduate: facility has a program for medical students only

4. Number of admission-prevalent CDI LabID events (identified within the first 3 days after admission to the facility, where date of admission = day 1).

5. Reporting of CDI labID events in Emergency Departments or 24-hour observation units.

6. Number of admissions to the facility.

7. Microbiological test method used to identify C. difficile (e.g., PCR for toxin, EIA assay for toxin, stool antigen, culture, other). The CDI testing algorithm of "NAAT plus EIA, if NAAT-positive" is currently receiving the "NAAT" level of risk adjustment under the 2017 NHSN protocol. Starting in 2018, the CDI testing algorithm of "NAAT plus EIA, if NAAT-positive" will be assigned the "EIA" level of risk adjustment.

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

Data from patients who are not assigned to an inpatient bed are excluded from the denominator counts, including outpatient clinics, 24-hour observation units, and emergency department visits. Inpatient rehab locations and inpatient psychiatric locations that have their own Centers for Medicare and Medicaid Services (CMS) Certification Number (CCN) are excluded. Additionally, data from NICU, SCN, babies in LDRP, well-baby nurseries, or well-baby clinics are excluded from the denominator count.

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.)*

Definition of inpatient - A patient who is located in an inpatient location for care and treatment at the time of the daily inpatient census count.

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.)*

The measure will not be stratified, as it is an overall facility-wide summary measure. Facility characteristics will be used for risk adjustment, described above in S9.

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:

Ratio

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 = Lower 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 Standardized Infection Ratio (SIR) for annual and quarterly data aggregation and analysis of CDI bacteremia LabID events is calculated for each healthcare facility for a specified time period. The SIR is an indirect standardization method for summarizing healthcare-associated infection (HAI) experience, including CDI bacteremia LabID events, in a single group of data or across any number of stratified groups of data. To produce the SIR:

1. Identify number of observed hospital-onset incident CDI LabID events for a given time period by adding the total number of observed events across the facility.

2. Calculate the number of predicted hospital-onset incident CDI LabID events for the facility using the methodology described. See attached table.

3. Divide the number of observed hospital-onset incident CDI LabID events (1 above) by the number of predicted hospital-onset incident CDI LabID events (2 above) to obtain the SIR.

4. Perform a mid-P Exact test to compare the SIR obtained in 3 above to the nominal value of 1. P-value and confidence interval will be calculated, which can be used to assess significance of SIR.

The Adjusted Ranking Metric (ARM) for annual data aggregation and analysis of HAI events, including CDI bacteremia LabID events, combines the method of indirect standardization used to calculate the unadjusted SIR described above with a Bayesian random effects hierarchical model to account for the potentially low precision and/or reliability inherent in the unadjusted SIR. A Bayesian posterior distribution constructed through Monte Carlo Markov Chain sampling is used to produce the adjusted numerator. The ARM enables more meaningful statistical differentiation between hospitals by accounting for differences in patient case-mix, exposure volume (e.g. patient days, central line-days, surgical procedure volume), and unmeasured factors that are not reflected in the unadjusted SIR and that cause variation between healthcare facilities. Accounting for these sources of variability enables better measure discrimination between facilities and leads to more reliable performance rankings. To produce the ARM:

1. Identify the number of hospital-onset incident CDI LabID events for the facility

2. Obtain the adjusted number of observed hospital-onset incident CDI LabID events for the facility using a Bayesian posterior distribution constructed through Monte Carlo Markov Chain sampling which results from a Bayesian random effects model.

3. Total these numbers for an observed number of hospital-onset incident CDI LabID events

4. Obtain the predicted number of hospital-onset incident CDI LabID events for the facility following the methodology provided (see attachment for final risk adjustment model).

5. Divide the total number of adjusted hospital-onset incident CDI LabID events (3 above) by the predicted number of hospital-onset incident CDI LabID events (4 above) to obtain the reliability-adjusted SIR

6. Perform a Poisson test to compare the SIR obtained in 5 above to the nominal value of 1. P-value and confidence interval will be calculated, which can be used to assess significance of SIR.

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.

No sampling methodology is used in calculating the metric.

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.

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

If other, please describe in S.18.

Electronic Health Data, Electronic Health Records, Other, Paper Medical Records

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.

NHSN Laboratory-identified MDRO or CDI Event Form and NHSN MDRO and CDI Prevention Process and Outcome Measures Monthly Monitoring Form

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

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

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

Facility, Other, Population : Regional and State

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

Emergency Department and Services, 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.)*

2. Validity – See attached Measure Testing Submission Form

NQF_CDI_ACH_attachment_2018_Final.docx,NQF_CDI_Testing_Final_July_27_Edit_per_NQF_Submitted_Aug_16.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.

No

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)

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: National Healthcare Safety Network	☐ other: Click here to describe

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*).

National Healthcare Safety Network

1.3. What are the dates of the data used in testing? Click here to enter date range

January 1- December 31, 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**

☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: Population: Regional and State	☒ other: Population: Regional and State

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*)

The standard population's hospital-onset CDI rates that were used in the SIR calculation came from facility-wide inpatient locations (FacWideIn) reporting CDI LabID events to NHSN from January 1 to December 31, 2015.This represented 3,613 reporting for 14,038 FacWideIn facility-quarters.

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*)

Yearly pooled NHSN organization Identification (Org ID)-level numerators ranged from 0 to 98 CDI LabID events (IQR 0 to 10), and the denominators ranged from 1 to 93,040 patient days (IQR 2,206 to 13, 858).

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.

All data followed the same exclusion rules: zero patient days, missing survey (i.e. risk-adjustment variables), outlier community onset prevalence rate.

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.

No patient-level sociodemographic variables are used in the measure and none were available for analysis.

_____

2a2. RELIABILITY TESTING

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

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

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

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

No additional systematic testing was conducted as the value of the measure as an indicator for differentiating good and poor performance and has been substantiated by its broad use for that purpose. This measure is widely used by healthcare facilities and state health departments to inform their CDI surveillance and prevention efforts, by prevention collaboratives to identify intervention opportunities and measure impact of interventions, and by the Centers for Medicare and Medicaid Services for the agency's public reporting and

payment programs. In our experience, questions and concerns about the validity of CDI definition and criteria are infrequent and typically reflect a misunderstanding of the definition and criteria.

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*)

No additional testing was conducted because the measure is widely used.

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*)

No additional testing was performed because the measure is widely used.

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?*)

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☐ **Performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

New empirical testing of the measure score has been conducted as the measure's value as an indicator for differentiating poor and good performance has been established through its wide use.

NHSN provides guidance to State Health Departments for conducting external validation of HAI data reported by facilities to NHSN within their jurisdiction. The validation process includes selection of sample of facilities and subsequently a sample of charts from the selected facilities which are reviewed by trained chart abstractors and tally against the data reported to NHSN. Case classification during the medical chart review and application of the protocol by the auditor is considered as the gold standard and compared with the facility determinations. Data accuracy measures assessed include sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity of CDI reporting is the correct identification of toxin positive specimens meeting CDI LabID criteria as CDI (true positive rate), whereas specificity is the correct identification of a positive toxin specimens not meeting CDI LabID criteria as "not CDI" (true negative rate). The positive and negative predictive values (PPV and NPV respectively) are the proportions of true positive and negative CDI's among all results that are reported by the facility during the time frame that are positive and negative, respectively. Total CDI mismatches included total number of CDIs missed and over-reported and an "error rate" was computed as a percent proportion of mismatches among total number of records reviewed.

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

CDI LabID State Validation snapshot:

State	Year of data validated	Records reviewed	sensitivity	specificity	PPV	NPV
Connecticut	2013	1085	93	99	99.9	75
Colorado	2015	359	95	100	100	80
Tennessee	2015	534	89.4	73.5	98	32
Utah	2016	394	92.5	100	100	42.9
New Mexico	2016	302	100	58.3	98.3	100
New York	2014	1787	89.4	100	100	42.8
Overall		**4461**	**94.9**	**93.7**	**99.4**	**58.7**

NHSN provides guidance to state health departments to assess the reporting accuracy of CDI data. Six state health departments have shared results of their CDI validation (Table above). We computed pooled mean estimates of reporting accuracy from the state validations and CDI reporting demonstrated high sensitivity (94.9%, range 89.4 – 100), high specificity (93.7%, range 58.3 -100), high PPV (99.4%, range 98.3-100) and low to moderate NPV (58.7%, range 32-100).

Pooled error rates in CDI case classification were computed. Error rate was computed as proportion of CDI cases incorrectly classified (missed and over-reported) among all the records reviewed. State CDI validations indicated a 5.9% pooled error rate of CDI reporting in NHSN.

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 measure has > 93% sensitivity/specificity.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

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*)

Neonatal intensive care units and well-baby nurseries are excluded from NHSN LabID event surveillance for CDI. Event numerators and patient day denominator counts for the measure do not include data from these locations. Furthermore, hospitals with an extremely high quarterly community-onset prevalence rate, defined as greater than 5 times the interquartile range of the national baseline, are excluded from the CDI SIR for that quarter.

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*)

Exclusions	orgIDs excluded	Facility-quarters excluded
ZERO or NULL patient days for quarter	44	304
Missing survey	8	52
Outlier Community Onset Prevalence Rate	3	104

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*)

CDI is not accurately identified in neonates less than 1 year of age due to colonization of the gut. Therefore, this age group is excluded from NHSN LabID Event surveillance for CDI. Furthermore, we concluded that significant adjustments to the parameter estimates would have occurred if we did not impose an exclusion criteria on the continuous variable in the model (community-onset prevalence rate). Due to likely data entry errors, hospitals reported prevalence rates ranging from 0 – 266.7 with an interdecile range of 0 to 0.0806.

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 2b5**.**

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 7 risk factors**

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

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

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

The risk model was conducted using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. Univariate models were fist constructed to evaluate the relationship between each risk factor and the CDI incidence rate. An assessment of outliers was performed on all continuous variables. Details of the final multivariate risk model are below:

Table 1. CDI in Acute Care Hospitals

Parameter	Parameter Estimate	Standard Error	P-value
Intercept	*-8.9463*	*0.0523*	*<0.0001*
Inpatient community-onset prevalence rate*	0.7339	0.0181	<0.0001
CDI test type+: EIA	-0.1579	0.0246	<0.0001
CDI test type+: NAAT	0.1307	0.0219	<0.0001
CDI test type+: OTHER	REFERENT	-	-
Medical school affiliation‡: Major, graduate, or undergraduate	0.0331	0.0111	0.0028
Medical school affiliation‡: Non-teaching	REFERENT	-	-
Number of ICU beds‡: ≥ 43	0.7465	0.0412	<0.0001
Number of ICU beds‡: 20- 42	0.7145	0.0395	<0.0001
Number of ICU beds‡: 10-19	0.6261	0.0396	<0.0001
Number of ICU beds‡: 5-9	0.4394	0.0420	<0.0001
Number of ICU beds‡: 0-4	REFERENT	-	-
Facility type: Oncology Hospital *(HOSP-ONC)*	1.2420	0.0765	<0.0001
Facility type: General Acute Care Hospital *(HOSP-GEN)*	0.3740	0.0342	<0.0001
Facility type: Other Specialty Hospital	REFERENT	-	-
Facility bed size‡	0.0003	0.0000	<0.0001
Reporting from ED or 24-hour observation unit^: YES	0.1119	0.0179	<0.0001
Reporting from ED or 24-hour observation unit^: NO	REFERENT	-	-

*Inpatient community-onset (CO) prevalence is calculated as the # of inpatient CO CDI events, divided by total admissions x 100 (i.e., cdif_admPrevCOCount /numCdifadms * 100). The prevalence rate for an entire quarter is used in the risk

+ CDI test type is reported on the FacWideIN MDRO denominator form on the 3rd month of each quarter and represents the testing method used by the laboratory to identify CDI. .

‡ Medical school affiliation, number of ICU beds, and facility bed size are taken from the NHSN Annual Hospital Survey.

Equation for # of predicted CDI events:

Exp [-8.9463
+ 0.7339 (CO prevalence rate**)**
- 0.1579 (CDI test type = EIA**)**
+0.1307 (CDI test type = NAAT**)**
+ 0.7465 (ICU beds ≥ 43**)**
+ 0.7145 (ICU beds: 20 – 42**)**
+ 0.6261 (ICU beds: 10-19**)**
+ 0.4394 (ICU beds: 5-9**)**
+1.2420 (Oncology hospital**)**
+ 0.3740 (General hospital**)**
+ 0.0003 (Total facility bed size**)**
+ 0.1119 (Reporting from ED or 24 hr. Obs**)**
+ 0.0331 (Teaching hospital**)] X CDI patient days**

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.

N/A

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?

Potential risk factors were selected based on availability in NHSN, literature review, and subject matter expert opinion. An expert panel from CDC DHQP was formed to identify potential risk factors in the beginning of model building process. First, all available facility-level variables from NHSN were presented to the expert panel. Facility and laboratory testing characteristics, as well as an indicator variable for cancer hospital were considered as potential risk factors.

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)**

Due to concerns about data entry burden and the paucity of evidence to support social risk factor data collection for risk adjustment purposes, social risk factors are not collected in NHSN for all patients in the patient population; therefore, these variables are not available in NHSN to be used for risk adjustment modeling.

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

Variables were eligible for entering the model at p-value=0.25 and retaining in the model at p-value=0.05 significant level. Factors were entered into a multivariate model using forward selection, based on the lowest Wald Chi-square value. Goodness of fit was assessed at each modeling step using the Akaike Information Criterion (AIC) statistics. The final model resulting from forward selection was confirmed via backwards elimination, in which each variable was sequentially removed based on the highest p-value.

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.**

N/A

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

Bootstrap sampling method was used to validate the models.

Model validation steps:

1. For each multiple logistic regression model, calculate the c-index as $C_{orginal}$.
2. Generate 100 bootstrap samples from the original dataset with the same number of records as the original sample size using sampling with replacement.
3. For each one of the new samples m=1, …,100, using the predictors of the logistic regression model from step 1 to fit the data with backward elimination approach and calculate the discrimination $as\ C_{boot}^{(m)}$. Note that the model we select from each of the m bootstrap samples could be different from the original model.
4. For each bootstrap sample, the original dataset is used for validation. For this step, the regression coefficients are fixed to their values from step 3 to determine the joint degree of over fitting from both selection and estimation. We obtain $C_{original}^{(m)}$ from this step.
5. For each one of the bootstrap samples, first we will calculate the optimism in the fit: $O^{(m)} = C_{boot}^{(m)} - C_{original}^{(m)}$. Then we obtain O by taking the average of $O^{(m)}$ from M bootstrap samples.
6. The optimism corrected performance of the original model is: $C_{adj} = C_{orginal} - O$. This value is a nearly unbiased estimate of the expected value of the optimism that would be obtained from external validation.

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

Model Validation Results for 2015 Rebaseline Model of CDI Events among Acute Care Hospitals

Parameter	Bootstrap Estimate	Empirical (2.5, 97.5) percentiles	N Converged
Intercept	-8.951	(-9.081, -8.823)	1000
CO_rate	0.7339	(0.6914, 0.7833)	1000
CDItest Type=EIA	-0.158	(-0.206, -0.108)	1000
CDItest Type=NAAT	0.1301	(0.0853, 0.1719)	1000

ICUBeds Q5	0.7478	(0.6506, 0.8504)	1000
ICUBeds Q4	0.7158	(0.623, 0.8204)	1000
ICUBeds Q3	0.6275	(0.5347, 0.7334)	1000
ICUBeds Q2	0.4404	(0.3368, 0.5518)	1000
Factype_3way=CANCER	1.2422	(1.0681, 1.4208)	1000
Factype_3way=GEN	0.3763	(0.2886, 0.4664)	1000
NumBeds	0.0003	(0.0002, 0.0004)	1000
ED_OBS	0.1129	(0.0723, 0.1531)	1000
Medschool	0.0333	(0.0105, 0.0544)	1000
Dispersion	0.1147	(0.1066, 0.1229)	1000

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

Likelihood Ratio Test, Akaike Information Criterion and dispersion-based adjusted R-squared

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

Likelihood Ratio Test, Akaike Information Criterion and dispersion-based adjusted R-squared

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

All Likelihood Ratio Tests for the best models indicated significant improvement as well as the lowest Akaike Information Criterion values and the greatest dispersion-based adjusted R-squared.

2b3.9. Results of Risk Stratification Analysis: N/A because bootstrap method was used.

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

The p-values for all variables in the final multivariate model were statistically significant, with most having a p-value < 0.0001. These variable are accounting for significant differences in risk of CDI between healthcare facilities. With the data reported to NHSN we have made full use of the available risk factor data to produce a series of prediction models for public reporting and pay for performance.

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*)

The multivariate regression model was confirmed and validated using bootstrap validation techniques.

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 CDI measure data are used to calculate an observed/predicted ratio, and ratios significantly higher than 1 are indicative of a quality concern that warrants full investigation and response.

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*)

Median 0.928

National Pooled mean 0.993

N= 3,634

Significantly higher than national SIR 427 (14%)

Significantly lower than national SIR 468 (15%)

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?*)

We see variation among facilities, and we can identify the facilities for which the summary measure warrants additional investigation.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

All facilities participating in NHSN and reporting LabID events to the MDRO module follow the same protocol for reporting events using similar laboratory and admission/discharge/transfer data sources.

Note*: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

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

All facilities participating in NHSN and reporting LabID events to the MDRO module follow the same protocol for reporting events using similar laboratory and admission/discharge/transfer data sources. The NHSN application provides "Alerts" to participating healthcare facilities in the event of missing data. In addition, CDC analysts conduct regular data quality checks and perform outreach to facilities regarding any missing or implausible data. Facilities that are not reporting data elements that are required by NHSN would not be eligible to receive an SIR.

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*)

Due to federal and state reporting requirements, as well as enforced business rules inside of the NHSN application, the majority of healthcare facilities are completing 100% of all required data entry, and thus minimal "missing" data exist.

Refer to table on 2b2.2. for exclusions

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

See above.

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.

Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry), Other

If other: LabID events and denominator data can be collected manually by trained hospital staff or via electronic data capture from hospital laboratory and ADT systems. The SIR is automatically calculated by the NHSN web application.

3b. Electronic Sources

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

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

ALL data elements are in defined fields in a combination of 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).

NHSN provides the option for facilities to collect the data electronically and download into NHSN. However, we leave the option for manual entry for facilities that are not equipped or ready to submit electronically.

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

Attachment:

3c. Data Collection Strategy

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

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

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

The NHSN Multidrug Resistant Organism and C. difficile Infection (MDRO/CDI) module has been available for facilities to use since 2009. The ability to perform facility-wide surveillance with a single denominator was introduced in 2010, reducing data collection burden on participating facilities. To further reduce case finding and data entry burden on facilities, LabID event reporting for C. difficile is now able to be performed electronically via NHSN's Clinical Document Architecture import function.

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)
Payment Program	Public Reporting Hospital Inpatient Quality Reporting Program (HIQR) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html LTCH Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/LTCH-Quality-Reporting/index.html Hospital-Acquired Condition Reduction Program (HACRP) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HAC/Hospital-Acquired-Conditions.html Public Health/Disease Surveillance National Healthcare Safety Network http://www.cdc.gov/nhsn/ Quality Improvement (Internal to the specific organization) Regulatory and Accreditation Programs https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html The Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) Program http://www.qualitynet.org/dcs/ContentServer?pagename=QnetPublic%2FPage%2FQnetTier2&cid=1228772356060 IRF Quality Reporting Program http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/IRF-Quality-Reporting-Program-Details.html

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

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

1) Name: Hospital Inpatient Quality Reporting Program (HIQR)
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: Nationwide, currently covers all acute care hospitals with ICUs (3201) and Non-IPPS (voluntary reporting): approx. 1,100*.
Level of measurement and setting: Facility-Level, acute inpatient hospital
2) Name: Prospective Payment System Exempt Cancer Hospital Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services

Purpose: To establish a quality reporting program for PPS-Exempt Cancer Hospital to improve health, improve care and lower cost (triple aims)of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: 11 Patient Prospective Payment Exempt Cancer Hospitals in 7 U.S. states with 19,203 average discharges each in FY 2012*.
Level of measurement and setting: Facility-Level, PPS-Exempt cancer hospital
3) Name: Inpatient Rehabilitation Facility (IRF) Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for IRFs to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients: All 50 U.S. States are included, 371,288 IRF discharges in 2011*.
Level of measurement and setting: Facility-Level, acute inpatient hospital
4) Name: Long Term Care Hospital (LTCH) Quality Reporting Program
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program for LTCHs to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: All 442 Medicare certified long-term care hospitals are required to participate to receive 100% of reimbursement money due. In 2012, this included 202,050 patient discharges*.
Level of measurement and setting: Facility-Level, LTAC inpatient
5) Name: Hospital Value-Based Purchasing
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: 2808 entities*
Level of measurement and setting: Facility-Level, acute inpatient hospital
6) Name: Hospital-Acquired Condition Reduction Program (HACRP)
Sponsor: Centers for Medicare and Medicaid Services
Purpose: To establish a quality reporting program to improve health, improve care and lower cost (triple aims) of Medicare beneficiaries.
Geographic area and number and percentage of accountable entities and patients included: 3,216 entities*
Level of measurement and setting: Facility-Level, acute inpatient hospital
*provided by Centers for Medicare and Medicaid Services

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?*)
N/A

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.

Numerous training materials have been created in order to assist users with the proper understanding and interpretation of this measure. Several webinars and written training materials have been provided. Annual in-

person trainings are held to discuss the SIR calculations, risk adjustment, and proper interpretation. Training materials are publicly available online as references for healthcare facilities and corporations, state health departments, and quality improvement organizations. NHSN users can run monthly analysis reports within NHSN to view their SIR data. On an annual basis, NHSN publishes national and state-level SIRs in the National and State Healthcare-associated Infection (HAI) Progress Report.

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.

SIR results are available to NHSN users at any time, based on their current data entry. Data provided within the analysis report includes numerator, denominator, SIR, p-value, and 95% confidence interval. Educational materials are available on the NHSN website that explain each data element.

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 on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, infection prevention and quality improvement staff, and other personnel. Based on user feedback through NHSN help desk, NHSN updated protocol to use final lab test result when performing a multistep testing methodology for CD identification. The final result of the last test finding which is placed onto the patient medical record will determine if the CDI laboratory assay definition is met, enabling use of test result which better reflects clinical determination.

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

Feedback from hospitals and states: Based on results from a polling survey, hospitals have indicated that they are running SIR analysis reports within NHSN on a monthly basis, and that they use SIRs for infection prevention activities in their hospital. State health departments are using the SIR for public reporting purposes and to help target facilities for additional prevention. Feedback regarding the extent of risk adjustment and the limitations.

was received via email and articles sited below:

Marra, A., Edmond, M., Ford, B., Herwaldt, L., Algwizani, A., & Diekema, D. (2018). Impact of 2018 Changes in National Healthcare Safety Network Surveillance for Clostridium difficile Laboratory-Identified Event Reporting. Infection Control & Hospital Epidemiology, 39(7), 886-888. doi:10.1017/ice.2018.86

Marra, AR, Ford, BA, Herwaldt, LA, Algwizani, AR, and Diekema, DJ. “Failure of Risk-Adjustment by Test Method for C. difficile Laboratory-Identified Event Reporting” Infection Control & Hospital Epidemiology 2016:1–3

Marra AR et al contend in their article published in the July 2018 issue of Infection Control and Hospital Epidemiology that the NHSN CDI LabID event healthcare quality performance measure does not adequately risk adjust for CDI test method (page 886). However, the authors are not clear about the appropriate litmus test for judging the adequacy of risk adjustment, and their critique of the NHSN risk adjustment is based on faulty premises and a flawed analytic strategy. They apply an unfounded method and deliver a misinformed critique. They used CDI data from a single hospital to draw conclusions about model performance across all hospitals that submit CDI LabID event data to NHSN. Specifically, they expected that with the NHSN CDI risk adjustment, identical CDI SIRs would be produced when either the NAAT or EIA testing algorithm was used at the hospital and reported to NHSN. Identical SIRs would occur only if their hospital adjusted CDI incidence was precisely at the mean. In other words, the NHSN CDI model--which accounts for other factors in addition to test type--shows that there is a 33.5% increase in CDI incidence comparing NAAT to EIA. This single hospital reported a 180% increase comparing NAAT to EIA, which clearly exceeds the mean. Additionally, this hospital’s CDI incidence rate using NAAT differs by a much wider margin from the relative national mean compared with the CDI incidence using EIA, a point that further illustrates why their SIRs using NAAT versus EIA are different. In effect, they propose that risk adjustment for the performance measure should be standardized to their CDI

experience rather than standardized to the mean value calculated using CDI data submitted by all hospitals to NHSN (3,613 acute care hospitals in 2015 reported CDI data that NHSN used to develop its CDI risk adjustment). The latter approach, which NHSN used to develop its risk adjustment for the performance measure, is standard and appropriate. Their analysis and proposal are methodologically unsound, and as such their input does not provide useful guidance for NHSN's risk adjustment methodology.

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

Feedback from consumers, media, policy, etc. on measure performance and implementation is obtained via email to the NHSN helpdesk email system. Feedback is provided to us by hospital staff, physicians, epidemiologists, statisticians, state and local health department staff, quality improvement staff, infection prevention and other personnel. See 4.a2.2.1.

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.

Feedback from all stakeholders is considered when developing and implementing the SIR. Different risk factor variables were analyzed for potential inclusion in the statistical model due to input from users. Additional training formats, such as live chats and "quick learn" videos, were created in order to address different training environment that best meet the needs of our audience. We have also provided live demonstrations to users showing how to generate their SIRs in NHSN based on earlier feedback we had received. See 4.a2.2.1.

Marra AR et al analysis and proposal are methodologically unsound, and does not provide useful guidance for NHSN's risk adjustment methodology.

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.

By the end of 2014 there had been an 8% decline in the SIR from baseline and between 2015 and 2016 there was another 8% decline using the 2011 baseline and 7% decline using the 2015 baseline. The most recent pace of progress needs to remain steady or increase to meet national prevention goals for hospital-onset CDI in 2020.Crude rates of healthcare-associated CDI are decreasing, which largely reflects declines in nursing home-onset infections, along with some declines in hospital-onset CDI.

https://www.cdc.gov/hai/surveillance/data-reports/data-summary-assessing-progress.html

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.

Laboratory and other clinical data must be reviewed to determine if the patient meets the criteria for a LabID event. It is possible that medical record reviewers will miss positive cultures or important dates that would indicate that a LabID event should be recorded. Similarly, reviewers might miss data in the medical record that would indicate a positive culture should not result in a

LabID event. It is also possible that data abstractors could intentionally underreport LabID events.

Business logic is built into the NHSN application to minimize incorrect entry of LabID events. Additionally, agencies including state health departments and others have indicated interest in performing validation of LabID event surveillance as they have for other healthcare-associated infections, such as central line-associated bloodstream infections.

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

N/A

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.

No

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

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?

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

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.)

N/A

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:** CDI_appendix_Final.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Disease Control and Prevention

Co.2 Point of Contact: Daniel, Pollock, MD, dpollock@cdc.gov, 404-639-4237-

Co.3 Measure Developer if different from Measure Steward: Centers for Disease Control and Prevention

Co.4 Point of Contact: Daniel, Pollock, MD, dpollock@cdc.gov, 404-639-4237-

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.

None

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2009

Ad.3 Month and Year of most recent revision: 07, 2018

Ad.4 What is your frequency for review/update of this measure? Annually and as needed

Ad.5 When is the next scheduled review/update for this measure? 07, 2019

Ad.6 Copyright statement: all CDC documents are public record therefore there is no copyright.

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments:

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

Brief Measure Information

NQF #: 3450 (previously NQF# 0206)

Measure Title: Practice Environment Scale - Nursing Work Index (PES-NWI) (composite and five subscales)

Measure Steward: University of Pennsylvania, Center for Health Outcomes and Policy Research

Brief Description of Measure: Practice Environment Scale-Nursing Work Index (PES-NWI) is a survey-based measure of the nursing practice environment completed by staff registered nurses; includes mean scores on index subscales and a composite mean of all subscale scores.

Developer Rationale: The dissemination of the PES-NWI nationally and internationally assures that nurses' practice environments will be measured in consistent fashion across different health systems to develop evidence guiding policy and management decisions. The benefit of using the PES-NWI measure for health care organizations is that organizations provide better quality patient care through improved work environments.

Numerator Statement: Continuous Variable Statement: For surveys completed by Registered Nurses (RN):

12a) Mean score on a composite of all subscale scores

12b) Mean score on Nurse Participation in Hospital Affairs (survey item numbers 5, 6, 11, 15, 17, 21, 23, 27, 28)

12c) Mean score on Nursing Foundations for Quality of Care (survey item numbers 4, 14, 18, 19, 22, 25, 26, 29, 30, 31)

12d) Mean score on Nurse Manager Ability, Leadership, and Support of Nurses (survey item numbers 3, 7, 10, 13, 20)

12e) Mean score on Staffing and Resource Adequacy (survey item numbers 1, 8, 9, 12)

12f) Mean score on Collegial Nurse-Physician Relations (survey item numbers 2, 16, 24)

12g) Three category variable indicating favorable, mixed, or unfavorable practice environments: favorable = four or more subscale means exceed 2.5; mixed = two or three subscale means exceed 2.5; unfavorable = zero or one subscales exceed 2.5.

Denominator Statement: Staff RNs

Denominator Exclusions: Not applicable

Measure Type: Structure

Data Source: Instrument-Based Data

Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: August 5, 2009 **Most Recent Endorsement Date:** December 14, 2012

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 *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- The developer provides a summary of several systematic literature reviews, including at least one (pre-publication) review and meta-analysis of the evidence connecting hospital nurses' work environments to patient outcomes.
- The developers summarize the results of that review, noting that better work environments were associated with lower odds of negative outcomes and higher odds of positive outcomes.

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:

Exception to evidence

NA

Questions for the Committee:

If the developer provided updated evidence for this measure:

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

Guidance from the Evidence Algorithm

Process measure with systematic review (Box 3) → Summary of the QQC provided (Box 4) → Systematic review concludes moderate quality evidence.

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

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.

- Performance data from the National Database of Nursing Quality Indicators is provided, covering the years 2013-2017.
- The developer notes that the sample hospitals exhibited the full range of possible scores (1.00 to 4.00), with standard deviations on the composite measure ranging from 0.29 to 0.31.

Disparities

NA

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

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

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

1a. Evidence
Comments:

**Some new evidence was provided that is directly related to the measure.

**The measure is supported by evidence that positive nursing working environment help improve quality of care and patient outcomes. The evidence has mostly been consistent over the years.

**A structure metric, literature associates better work environments with lower odds of negative outcomes and higher odds of positive outcomes.

**No concerns.

**Poor.

**Evidence that better work environments provide better outcomes is updated, directionally the same and stronger.

**Previous endorsements in 2009 and 2012 with updated evidence.

**Moderate - item wording is not consistently grammatical.

**Acceptable.

**Strong evidence.

**3450 (previously NQF# 0206); Structure Measure. 2017 and 2018 Systematic literature reviews. 46 articles from 28 countries. • Primary data collection occurred in 25 of the studies. The remaining 22 studies analyzed secondary data with the earliest reporting year of collection occurring in 1999 and the latest in 2014.

1b. Performance Gap
Comments:

**A performance gap remains. The issue of the work environment for nurses is a critical one, particularly in light of ongoing legislative involvement in nursing ratios.

** Performance gap is clearly shown with performance data.

**Moderate, performance gap.

** Higher scores assoc with better outcomes etc.

** Still subjective reporting.

** Full range of reporting from 1-4 by sample hospitals.

** Sample hospitals exhibited full range of possible scores 1-4.

** High.

** Variation exists; no info re "disparities".

** Variability noted.

** In a study by the measure developer, Lake, from 794 hospitals in 4 states, the sample hospitals exhibited the full range of possible scores: 1.00 to 4.00. The average hospital-level subscale scores ranged from 2.50 to 2.84, with SDs ranging from .29 to .37. The descriptive statistics calculated from all community hospitals in four states demonstrate lower average scores than the Joint Commission pilot hospitals as well as much greater variation across hospitals, suggesting that Joint Commission accredited hospitals have better nursing environments than all hospitals (in these 4 states and perhaps throughout the U.S.) and indicating the capacity of the PES-NWI measure to provide evidence of significant and meaningful differences in practice environmental performance across providers. Maintenance of Endorsement (October 2018): Descriptive statistics from the National Database of Nursing Quality Indicators revealed performance gaps.

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.

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Evaluation of Reliability and Validity (and composite construction, if applicable):

Scientific Methods Panel Votes: Measure passes

- Reliability: H-3, M-1, L-0, I-1
- Validity: H-3, M-1, L-0, I-1

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- Reliability was conducted at the data element and measure score level.
- Data element
    - Conducted by computing Cronbach's alpha.
    - Results: Provided overall summary of results based on 46 articles reviewed by Swiger et al (2017).
        - 37 articles reported Cronbach's alphas; coefficients ranged from .71 – .96, with the exception of one .67, and one .53 in a small sample size.
- Measure score
    - Conducted by assessing inter-rater reliability, which focuses on whether nurses give consistent responses within a hospital or nursing unit, as compared to across hospitals or nursing units in a sample. Performance measure score reliability is assessed using the intraclass correlation (ICC) (1,*k*),
    - Results: based on 14 articles below and the 2015 National Database of Nursing Quality Indicators nurse survey data:

Reference	# organizational units (hospitals or nursing units)	# nurses	ICC (1,k) statistics reported or summarized
Lake (2002)	16 magnet hospitals proportionate by regions of the country	1,610	.88 to .97
Lake et al (2006)	156 adult community hospitals in Pennsylvania	10,962	.67 to .82
Clarke (2007)	188 Pennsylvania general acute care hospitals	11,512	.70 to .90
Flynn et al (2010)	63 Medicare and Medicaid certified nursing homes in New Jersey	897	Composite: .68 Subscales range: .55 to .75
Brooks-Carthon et al (2011)	429 hospitals across four states (Florida, Pennsylvania, New Jersey and California)	98,000	Subscales range: .73 to .90
McHugh et al (2012)	396 adult, non-federal acute care hospitals across four states (CA, FL, NJ, PA)	16,241	.61
Kelly et al (2013)	320 hospitals across four states (CA, FL, NJ, PA)	3,217	.69
McHugh et al (2013)	564 Magnet and non-Magnet hospitals across four states (CA, FL, NJ, PA)	100,000	.81
Kelly et al (2014)	303 adult care hospitals across four states (CA, FL, NJ, PA)	55,159	.71
McHugh et al (2014)	534 hospitals across four states (CA, FL, NJ, PA)	26,005	.85
Carthon et al (2015)	419 acute care hospitals across three states (CA, FL, NJ, PA)	20,605	.74 to .91
Ma et al (2015)	373 hospitals from 44 states	33,845	Ranged from .80 to .87
Lake et al (2016)	171 hospitals across four states (CA, FL, NJ, PA)	1,247	4 subscales >.60; 5th = .58
Swiger et al (2018)	45 acute care units in 10 Army hospitals	180	ICC (1,*k*) reported as satisfactory

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

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

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

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

Combined Methods Panel Evaluation: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

*Note: Completed by multiple Scientific Methods Panel members and therefore multiple responses provided in checkboxes.

Measure Number: 3450

Measure Title: Practice Environment Scale –-Nursing Work Index (PES-NWI)

Type of measure:

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

Data Source:

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

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

   PANEL MEMBER 2: The specs clearly delineate the methodology, numerator, and denominator. Specifically, a survey for nurses employed at a facility to assess the practice environment. It is comprised of five subscales, composed of three to ten questions per subscale, and a composite score of all of the

subscales. Means are calculated for each of the subscales and the composite is the mean of the subscale scores. There is also a three-category variable that portrays a favorable, mixed, or unfavorable environment, based on the number of subscales that exceed 2.5. This practice environment assessment has be utilized across many types of facilities for almost thirty years and is intuitively friendly to nurses of all types.

2. **Briefly summarize any concerns about the measure specifications.**
    - PANEL MEMBER 2: None.
    - PANEL MEMBER 4: None.
    - PANEL MEMBER 1: My primary concern with the measure specification relates to sampling instructions for the survey. The authors indicate a random sample of all RNs is used with a minimum of 30 responses as adequate to characterize the environment of the hospital. They reference the Joint Commission for this rule, but provide no evidence as it relates to the variation with the measure or the size of the hospital to justify this N. Although it's a bit odd to think of sampling instructions as part of the specification, the representativeness of the respondents seems to the at the heart of this type of process measure.
    - PANEL MEMBER 5: Cronbach Alpha for scales high enough. Some earlier literature that tried to reproduce the factor structure did not do so, but scales have conceptual validity, and Cronbach alpha sufficiently high to justify continued use.
    - PANEL MEMBER 3: Minor point – mean score is reported as the numerator. The sum of the scores is actually the numerator.
        - No data dictionary submitted – should be one, though simple and based on the instrument.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

    - PANEL MEMBER 2: Reliability testing at the data element level was conducted by computing Cronbach's alpha. It has been reported to range between 0.71 and 0.96 in cited a review by Wasrshawsky & Havens of 37 samples of 13 – 72, 889 nurses from 1998 to 2010, and one by Swiger of 46 samples of 133 to 33, 845 nurses form 2010 to 2016. Subsequent to 2016, 34 other references are provided with Cronbach's alpha ranging from 0.60 to 0.91 for the subscales, and between 0.80 and 0.91 for the composite. Reliability of the performance measure score was assessed by intraclass correlation coefficient. In the submission for maintenance in 2018 and using 2015 NDNQI data, the ICC's for the subscales were 0.936 to 0.973 over 451 hospitals and about 157,000 nurses. The ICC for the composite measure was 0.966. In a series of reviews from 2002 to 2017, the ICC's for the subscales and the composite are provided.

- PANEL MEMBER 4: Am concerned that the ICC's cited are too high (i.e. the denominator did not include all 3 sources of error – between hospitals, within hospitals across nurses, and with nurse across items in back scale).
- PANEL MEMBER 5: Cronbach assess subscale structure
    - ICC to assess whether nurses within units or hospitals (not always clear how done) reported similar scores.
- PANEL MEMBER 3: Chonbach's alpha at the nurse level
    - ICC at the hospital level
- PANEL MEMBER 1: For reliability at the item level the authors mention the Cronbach's alpha from numerous studies that used the scale. This suggests internal consistency at the scale level.
    - For reliability of the score, the authors looked at the consistency of RN scores within hospitals or units as compared to scores between hospitals or units. They use an ICC for this with a target of 0.60 or higher.

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

   Submission document: Testing attachment, section 2a2.3

   - PANEL MEMBER 2: This is a time-tested survey which has demonstrated high reliability at the data level and at the performance score level consistently. It has been replicated many times across many facility settings and across many nursing subsets.
   - PANEL MEMBER 5: Cronbach alpha assessments of scale structure sufficient to establish score.
       - ICC from reported studies vary from moderate (0.5-.75) to good or better.
       - Key ICC data for renewed endorsement is from National Database of Nursing Quality Indicators and reported ICCs for scale components and overall composite exceed 0.9. Ranges of results and histograms presented but no formal signal to noise analysis comparing within unit variance to between unit variance.
       - Would have liked to see some discussion of response rates in organizations using survey, such as for NDNQI reporting.
   - PANEL MEMBER 4: See above.
   - PANEL MEMBER 3: Chronbach's alpha 0.71 – 0.96
       - ICC 0.61 to 0.97
   - PANEL MEMBER 1: The authors find the subscales internally consistent with a range of Cronback's alphas from 0.71 to 0.96 form 37 studies using the survey tool. In terms of the score, the authors report an ICC of 0.966 for the overall measure with sub-scale scores ranging from.936 to .973 from a sample of over 400 hospitals and more than 150,000 nurses. The scores are more mixed for individual studies using the tool.

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 2: I have no concerns with this consistently highly reliable survey and the assessment methodology has been replicated across many settings and populations.
- PANEL MEMBER 5: I'm split between moderate and high reliability rating. ICC results from NDNQI are strong, but results from other reported research more mixed. Would have like to see signal to noise assessment beyond ICC to better understand the extent of variance across units compared to within unit variance.
- PANEL MEMBER 4: See #6.
- PANEL MEMBER 1: Although the ICC scores varied across studies with some fair and good scores, the preponderance of evidence seems to be strong within unit correlation.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

- **PANEL MEMBER 1**: No concerns with exclusion of shift and visiting nurses.
- **PANEL MEMBER 2**: There are no measure exclusions.
- **PANEL MEMBER 3:** No exclusions - NA
- **PANEL MEMBER 4**: None.
- **PANEL MEMBER 5:** None. If individuals skip items, excluded items not taken into account for average score, but data on missing data suggests little skipping.

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

    Submission document: Testing attachment, section 2b4.

- **PANEL MEMBER 2:** The two prior reviews cited above by Warshawsky & Havens, and by Swiger, demonstrated meaningful variation identified by both the composite performance scores and the subscores. No concerns.
- **PANEL MEMBER 3**: Descriptive statistics only – no statistical testing regarding meaningful differences.
    - Unclear how results will be used to improve performance or meaning of low/high performance.
- **PANEL MEMBER 5**: Scores range from 1-4, four point scale. Average reported range for individual scales and composite roughly 1 to 1.5 points. This seems to be a reasonable range for meaningful differences, particularly given reported correlation with other measures of nursing performance.

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

- **PANEL MEMBER 1:** The authors indicate there is a web-based and paper-based option for this tool. They do not seem to do much comparison between the two -- most likely the questions are identical. Having both options is important for clinical settings with RNs do not have unique email addresses. The authors do mention a push to have the tool be all electronic at some point in the future.
- **PANEL MEMBER 2:** The survey is the only data source. No concerns.
- **PANEL MEMBER 3:** NA
- **PANEL MEMBER 5:** Main concern is that survey is used in alternative settings (research studies with investigators not attached to nurses hospitals; individual hospital or unit sponsored surveys). No systematic analysis of variations in response rates, ICC or scores when survey administered in different contexts.

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

Submission document: Testing attachment, section 2b6.

- **PANEL MEMBER 1:** As mentioned earlier, the survey response rate is a concern for me. I'd like to be see more about the response rates for the various studies they references and the characteristics of non-responders. The authors report very low. rates of missing items within the survey.
- **PANEL MEMBER 2:** Missing data is reported at both the data element level and at the facility level, the former representing less than 1% of the respondents, and the latter demonstrating that 90% of the facilities have less than 4% of their respondents with missing data. The conclusion reached is that this represents random variation and is minimal.
- **PANEL MEMBER 3**: Mention of 'missing data calculated', but no methodology details.
- **PANEL MEMBER 5:** Skips of questions appear to be low
    - Typical response rates and bias in response not discussed.

16. **Risk Adjustment**

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

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

☒ Yes ☒ No ☐ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☒ Not applicable

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☒ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☒ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☒ No

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

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

- **PANEL MEMBER 2:** The measure developers present a rational argument that since the survey is completed by nurses at facilities, there is no reason to suspect interference by any demographic variables that would impact the results.
- **PANEL MEMBER 4:** No rationale provided.
- **PANEL MEMBER 5:** No risk adjustment. Patient level risk adjustment not relevant.
    - Discussion of whether to risk adjust based on nurse characteristics (e.g., tenure, education, age) is conceptual in rejecting adjustment, but no systematic analysis of variance with units in scores along any of these dimensions.

VALIDITY: TESTING

17. **Validity testing level: ☒ Measure score ☐ Data element ☐ Both**
18. **Method of establishing validity of the measure score:**
    - ☐ **Face validity**
    - ☒ **Empirical validity testing of the measure score**
    - ☐ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

- **PANEL MEMBER 1**: The authors focus on results from two review articles, looking at construct, concurrent and predictive validity. One of the primary comparisons is between magnet and non-magnet hospitals, based on the assumption that work environment should be better in magnet hospitals. They go on to show negative and positive associations with other outcome measures such as mortality, infections and falls.
- **PANEL MEMBER 2:** The measure score was tested for validity by the statistical association between the measure and the hypothesized related constructs. Developed in 2002 through factor analysis of 1986 survey data from nurses in 16 original magnet hospitals, it was then retested and confirmed in 1999 data from over 11,636 nurses in Pennsylvania. The five subscales are combined into a composite measure of the practice environment, either a continuous variable, or a three-category variable indicating favorable, mixed, or unfavorable practice environments. The appropriate references are provided. Since development, the measure has undergone testing with correlation coefficients, ANOVA, t-tests, and regression coefficient analysis.
    - In Warsharwsky & Havens 2011 article, the majority of the 37 studies associated the results with nurse outcomes (job satisfaction, intent to leave, burnout, and work engagement), patient outcomes (patient satisfaction, medication errors), and organizational outcomes (safety climate and morale). This was repeated in the Swiger study. Further analysis compared non-magnet versus magnet facilities, an indicator of quality nursing and organizational excellence. The results were not only correlative but discriminant.
- **PANEL MEMBER 3:** Discriminant validity tested – Magnet vs non-Magnet facilities.
- **PANEL MEMBER 4:** The developer cites data from 2006-2008 in published literature. No data were provided for the "3 category" summary measure.
- **PANEL MEMBER 5:** Principal approach is to demonstrate concurrent validity, with significant correlation of score with patient outcomes, nurse reported satisfaction and burnout, etc., and patient assessments of quality of care using HCAHPS.

20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

- **PANEL MEMBER 1:** With the magnet hospital analysis, the authors show the range of scores for magnet and non-magnet hospitals. The ranges do overlap, even though the magnet hospitals are always towards the upper end of the range. On the concurrent validity, the long list of measures is impressive, but it would have been more convincing to see a smaller number of measures with a strong theoretical relationship to nurse working environment. The authors mention predictive validity, but it is difficult to tease this out with the large volume of results. Perhaps that was the HCAPHS analysis.
- **PANEL MEMBER 2:** This survey has been validity testing pre-development and on numerous post-implementation occasions and has consistently shown validity to other commonly accepted measures of nursing outcomes, patient outcomes, and organizational outcomes. These are summarize in Table 21b.3b. The evidence is strong and repetitive over time.
- **PANEL MEMBER 3:** Adequate – supported with literature and not tested by the measure steward.
- **PANEL MEMBER 4:** Although the range in hospital-level scores is provided for each subscale, the means and standard errors are not; neither are the frequencies for the "3 category" summary measure. They do provide hospital-level frequency distributions from 2015 data for each subscale from 452 hospitals.
- **PANEL MEMBER 5:** Substantial, extensive research base demonstrating association of NWI scores and patient outcomes, patient assessment of quality of care, and nurse satisfaction and burnout.

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

    ☒ **Yes**

    ☐ **No**

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

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

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

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

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

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

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

- **PANEL MEMBER 1:** There is a tremendous amount of evidence that the work environment scale is related to many different process and outcome measures. It would be helpful to have more of a theoretical justification for which outcomes are driven by work environment rather than correlated with work environment.
- **PANEL MEMBER 2:** As stated in Question 22, this is a well-tested, consistent, and reproducible rating of validity.
- **PANEL MEMBER 3:** No concerns
- **PANEL MEMBER 5:** High concurrent validity and face validity of individual items. Subscale structure is internally consistent per Cronbach's alpha but would like additional discussion of literature trying to reconstruct factor analysis results that created the subscales.

ADDITIONAL RECOMMENDATIONS

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

- **PANEL MEMBER 1:** My primary recommendation is a power analysis around sample size. It is not clear to me why an N of 30 is sufficient for complex, large hospitals.
- **PANEL MEMBER 4:** This is a widely used measure with considerable evidence of a link between nurses' working environment and quality of care. However, the ICC's cited are greater than the internal consistency coefficients which is highly unlikely given the within nurse across item variation. It should be included in the denominator in assessing within hospital variation. A regression spline showing hospital level means and standard errors would help.
- **PANEL MEMBER 5:** Overall, judge validity and reliability to be acceptable.
    - Would have liked more assessment of within unit vs between unit variance, and some discussion of factor analysis basis for scales.

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

2a1. Reliability – Specifications
Comments:

** No concerns.

** I would agree with one of the reviewers from the Scientific Methods Panel that it seems odd to use the minimum of 30 responses as adequate to characterize the hospital nursing working environment regardless the size of a hospital. I too would like to see a statistic justification, which is not provided by the sponsors.

**Specifications provided.

**No concerns.

**Too subjective.

**Clear definitions.

**No comment.

**High demonstrated by cronbachs alpha and kappa.

**Acceptable.

** No concerns.

**Of the 46 articles reviewed in Swiger et al (2017) published from 2010 to 2016, 37 reported Cronbach's alphas; coefficients ranged from .71 – .96, with the exception of one .67, and one .53 in a small sample size. These results support the coherence of the different subscales and the composite. Additional internal consistency reliability data were displayed.

2a2. Reliability – Testing

Comments:

**None.

**In general, no. But, I would agree with one of the reviewers from the Scientific Methods Panel that it seems odd to use the minimum of 30 responses as adequate to characterize the hospital nursing working environment regardless the size of a hospital. I too would like to see a statistic justification, which is not provided by the sponsors.

**Moderate, data element and inter-rater reliability tested, Scientific Methods Panel satisfied with results.

**No.

**Is there inter-rater reliability?

**No.

**None.

**No.

**Small - some ICCs only moderate.

**As above.

**None.

**None; Updated intraclass correlation cofficients 2018, all about 0.60.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences

Comments:

**No.

**No.

**Moderate, empirical validity testing

**No.

**No.

**None.

**No.

**Moderate.

**No.

**None.

**The method of validity testing was by statistical association between the measure and hypothesized related constructs, to demonstrate construct, concurrent, and predictive validity.

**No concerns.

**No.

**No concerns.

**No.

**No.

**Random sampling could give skewed results.

**No.

**Some low response rates, no gold standard validation variables.

**No concerns.

**None.

**None.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**No concerns.

**Does not seem to apply.

**None.

**No risk adjustment-process measure.

**I think there needs to be a trend and work environment at the time analysis.

**No concerns.

**Appropriate.

**None.

**None.

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.

- Data for the measure are generated through the Practice Environment Scale-Nursing Work Index (PES-NWI) Survey
- The survey can be collected through electronic survey software or via paper
- The developers suggest that a minimum of 30 responses per year are required to establish a m,inimum sample, and recommend that hospitals survey all eligible nurses.

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?*
- *Is the data collection strategy ready to be put into operational use?*

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

** No concerns about feasibility.

** None I can see.

** Moderate, electronic or paper survey.

** No concerns.

** Very dependent upon hospital.

** PES-NWI survey electronic or paper.

** Agree with moderate feasibility prelim rating.

** Moderate - response rates a potential problems, 30 responses per year seems too few.

** No concerns.

**None.

** The record of use of the measure by the NDNQI, the VA, and the military hospital systems demonstrates that there are minimal difficulties regarding data collection, availability of data, missing data (which was documented in the Measure Testing Submission Form submitted on 7/31/18, timing and frequency of data collection, sampling, nurse confidentiality, time and cost of data collection, or any other feasibility/implementation issues.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

- The measure is used in a number of accountability programs, including public reporting of results in at least one state (Colorado).

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 [vetting]

- The developer notes that this measure is used in the National Database of Nursing Quality Indicators (NDNQI), the VA, and military hospitals, and that performance results are shared in reports and dashboards with hospital managers.
- The NDNQI publishes monographs and holds conferences to provide feedback and guidance to facilities collecting and reporting the measure.
- The developer suggests that the increasing trend in completion of the measure indicates that the measure is valued by participating facilities.
- With respect to feedback from measured entities and other users, the developer notes that typical feedback on the measure is that a reduction in length and testing for use in non-hospital settings is desired.

Additional Feedback:

- NA

***Questions for the Committee**:*

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

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

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

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

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

Improvement results

- The developer suggests hospitals that have achieved "magnet" recognition for meeting standards of excellence have improved performance by having better work environments as compared to non-magnet hospitals.
- Performance results provided in section 1b of the measure submission also show slight improvements in average performance year over year.

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 notes some inconsistency in reporting of subscales and composites across studies, and variation in the unit of analysis for reporting.
- Otherwise, the developer notes no unexpected findings.

Potential harms NA

Additional Feedback: NA

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

4a1. Use - Accountability and Transparency
Comments:
**Already in use in some areas.
** See next comments.
**High, publically reported in accountability programs.
** Not sure.
** I think this is more a co-factor I. analysis rather than a measure in itself.
**Currently used in accountability programs, feedback given
** Agree with prelim rating.
**Pass.
** Since orignial emdoresement & re-endorsement only used in 1 state for public reporting; is this a measure that really matters?
** No concerns - used in some public reporting.
** The record of use of the measure by the NDNQI, the VA, and the military hospital systems demonstrates that there are minimal difficulties regarding data collection, availability of data, missing data (which was documented in the Measure Testing Submission Form submitted on 7/31/18, timing and frequency of data collection, sampling, nurse confidentiality, time and cost of data collection, or any other feasibility/implementation issues.

4b1. Usability – Improvement
Comments:

**Understanding the views of nurses regarding the organization, management and workflow of care is critical to quality improvement.

** According to AHA, there are 6,210 hospitals in the US. But, the number of hospitals that responded to the survey was hovering around only several hundreds. I am wondering why there aren't more hospitals participating in the survey and how to get more hospitals to participate? The measure was initially endorsed in 2009, which has been nearly 10 years. In the integrative literature review by Lee and Scott (2018), they found that the effects of work environment on patient outcomes were inconsistent. Both this study and the one by Warshawsky and Havens (2011) recommended further study to clearly identify the relationship between the nurse work environment and patient safety outcomes, in particular longitudinal studies. So I would highly recommend that the future evidence should show data that demonstrate how the adopted measure in the past 10+ years has improved patient safety outcomes and reduced medical harm, not just associations. I also think that the survey misses an important nursing work environment indicator, which is a working environment that encourages nursing staff to speak up about patient safety concerns. According to AHRQ, Staff willingness to speak up when they are concerned about unsafe behaviors and conditions is a hallmark of safety culture in a healthcare work environment. As recently reported by Palatnik, (Speak up for patient safety, Nursing Critical Care: 2016 , 11), a study collected data from more than 1,700 healthcare employees, including 1,143 nurses. The participants in the study reported frequent observation of colleagues making mistakes, appearing critically incompetent, or taking dangerous shortcut. However, less than 1 in 10 spoke up about their concerns. The study shows that a work environment with ineffective communication often displays failure of staff to speak up when they know something is wrong that could potentially cause harm to the patient. So I think it is important for the sponsor to update the PES-NWI to include a measure on a hospital's safety culture environment that encourages staff speaking up.

**No unexpected harms.

**No concern.

** Okay.

** No harms.

**No harms identified by developer.

**Moderate.

**Not sure how it is really used; does it make a demonstrable difference to hospital or pt outcomes?

**No concerns.

** Public Reporting Maintenance of Endorsement (October 2018): The state of Colorado collects PES-NWI data from all hospitals with at least 100 beds every two years (odd years). These data are publicly reported. Professional Certification or Recognition Program American Nurses Credentialing Center Magnet Recognition Program. Quality Improvement (external benchmarking to organizations) National Database of Nursing Quality Indicators. National Database of Nursing Quality Indicators. Centers for Medicare & Medicaid Services (CMS) Hospital Inpatient Quality Reporting (Hospital IQR) program Structural Measure: Participation in a Systematic Clinical Database Registry for Nursing Sensitive Care. Many states have mandated collection and reporting of nursing-sensitive measures, for example: Colorado: The Colorado Hospital Report Card.

Criterion 5: Related and Competing Measures

Related or competing measures

NA

Harmonization

NA

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

5. Related and Competing

Comments

**No.

**I am not aware of any.

**None.

**No.

**Not sure.

**None.

**No.

**None.

**Not specifically conflicting.

**None.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

Public Comment

** The PES-NWI is a well-recognized, valid tool for measuring nurses work environments. Since a positive work environment is linked to patient safety, I strongly support NQF continued endorsement.

** The PES-NWI is a valid, reliable tool to measure the nurse work environment. In my research using the PES-NWI, I have found that it performs consistently in different samples in terms of having above an acceptable Cronbach alpha level. It has stood the test of time and is a globally used measure for the nurse work environment that has found to be associated with patient and nurse outcomes. I highly recommend continuing National Quality Forum endorsement for the PES-NWI.

** The PES-NWI is a well recognized, valid, and reliable instrument for the measurement of nurses work environments. The PES-NWI has been an important measure for describing differences in healthcare quality across numerous settings as well as linking variation in nurses practice environments with differences in patient outcomes. The PES-NWI is widely used by numerous organizations and researchers both nationally as well as internationally. It clearly meets each of the measurement criteria at a high level. I strongly support ongoing endorsement of the PES-NWI.

** The PES-NWI is a recognized instrument to measure various elements of the nursing practice environment. Numerous publications highlight the breath and depth of the variables in various practice settings and countries. Identified subscales specify characteristics of the measures.; Widely used and accepted, the research findings continue since first introduced in 2004. Comparisons of nursing practice environment scores and adverse events/outcomes support the need to use the instrument findings as part of nursing leaders strategic intitiatives to improve quality and safety of nursing care. The instrument is invaluable to employers, nursing leaders and ultimately, patients. I urge further strong support of this nursing practice measure.

** The PES-NWI remains a commonly used and reliable instrument with which to measure the nursing practice environment. The large body of reserch demostrating associations between PES-NWI scores and adverse events/outcomes underpins the value and utility of this instrument. Continued use and analysis of the relationships deonstrated, particularly with regard to the instrument subscales, provides nursing leaders with actional information to use when they aim to improve the quality and safety of nursing care via improvements in the nursing practice envionment. Please endorse this measure.

Support/Non-Support

- There have been no comments or support/non-support choices as of this date.

Brief Measure Information

NQF #: 3450

Corresponding Measures:

De.2. Measure Title: Practice Environment Scale - Nursing Work Index (PES-NWI) (composite and five subscales)

Co.1.1. Measure Steward: University of Pennsylvania, Center for Health Outcomes and Policy Research

De.3. Brief Description of Measure: Practice Environment Scale-Nursing Work Index (PES-NWI) is a survey-based measure of the nursing practice environment completed by staff registered nurses; includes mean scores on index subscales and a composite mean of all subscale scores.

1b.1. Developer Rationale: The dissemination of the PES-NWI nationally and internationally assures that nurses' practice environments will be measured in consistent fashion across different health systems to develop evidence guiding policy and management decisions. The benefit of using the PES-NWI measure for health care organizations is that organizations provide better quality patient care through improved work environments.

S.4. Numerator Statement: Continuous Variable Statement: For surveys completed by Registered Nurses (RN):

12a) Mean score on a composite of all subscale scores

12b) Mean score on Nurse Participation in Hospital Affairs (survey item numbers 5, 6, 11, 15, 17, 21, 23, 27, 28)

12c) Mean score on Nursing Foundations for Quality of Care (survey item numbers 4, 14, 18, 19, 22, 25, 26, 29, 30, 31)

12d) Mean score on Nurse Manager Ability, Leadership, and Support of Nurses (survey item numbers 3, 7, 10, 13, 20)

12e) Mean score on Staffing and Resource Adequacy (survey item numbers 1, 8, 9, 12)

12f) Mean score on Collegial Nurse-Physician Relations (survey item numbers 2, 16, 24)

12g) Three category variable indicating favorable, mixed, or unfavorable practice environments: favorable = four or more subscale means exceed 2.5; mixed = two or three subscale means exceed 2.5; unfavorable = zero or one subscales exceed 2.5.

S.6. Denominator Statement: Staff RNs

S.8. Denominator Exclusions: Not applicable

De.1. Measure Type: Structure

S.17. Data Source: Instrument-Based Data

S.20. Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

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? Maintenance of Endorsement (October 2018): N/A

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

0206_Evidence_CompositeMSF1.0_Data-636682914757252218.doc,NQF_evidence_attachment_Sep2017_3450_Nov_1_2018.docx

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

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

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3450

Measure Title: Practice Environment Scale of the Nursing Work Index

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 11/1/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☐ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

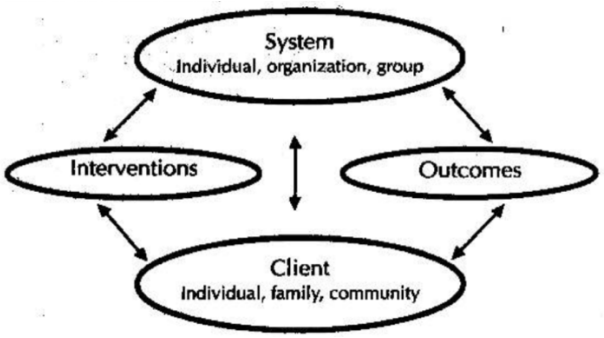
☐ Appropriate use measure: Click here to name what is being measured

☒ Structure: The Nursing Practice Environment

☐ Composite: Click here to name what is being measured

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

We place the nursing practice environment in the system characteristics within the Quality Health Outcomes Model (Mitchell et al. 1998), which postulates that interventions are mediated by system and client characteristics in influencing health outcomes.



Mitchell, P. H., Ferketich, S., Jennings, B. M., & American Academy of Nursing Expert Panel on Quality Health Care. (1998). Quality health outcomes model. Image: Journal of Nursing Scholarship, 30(1), 43-46.

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

This measure is derived from surveys of staff nurses. We theorize that there are two target populations that value the measured structure: nurses themselves and the patients they care for. This theory is based on the following factors:

- **The nursing workforce is the largest group of caregivers in all health care settings.**
- **All health care settings have nursing practice environments that may or may not support professional nursing practice. Therefore, practice environments, through their support of professional nursing practice, affect large numbers of health care providers and patients, affect the use of resources, and are the context of nursing care for patients facing all causes of morbidity and mortality and for all health care procedures.**

The Practice Environment Scale of the Nursing Work Index was derived from a larger instrument designed to measure organizational characteristics that influence nurse job satisfaction and perceived productivity (Kramer & Hafner, 1989). Therefore, the instrument content is valued and meaningful to nurses.

Kramer, M., & Hafner, L. (1989). Shared values: Impact on staff nurse job satisfaction and perceived productivity. Nursing Research, 38(3), 172-177.

Evidence regarding patient's perceptions of the quality of health services in general and hospital care in particular shows that patients acknowledge the nurse's contributions to quality of care, particularly the nurse's role in communication, and are aware of work environment factors, including nurse and physician collaboration, and the adequacy of nurse staffing (Sofaer et al. 2005; Sofaer & Firminger, 2005).

Sofaer, S., Crofton, C., Goldstein, E., Hoy, E., & Crabb, J. (2005). What do consumers want to know about the quality of care in hospitals? Health services research, 40(6p2), 2018-2036.

Sofaer, S., & Firminger, K. (2005). Patient perceptions of the quality of health services. Annual review of public health, 26.

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

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

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☒ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **The Practice Environment Scale of the Nursing Work Index: An updated review and recommendations for use** • **Pauline A. Swiger, Patricia A. Patrician, Rebecca S. (Susie) Miltner, Dheeraj Raju, Sara Breckenridge-Sproat, Lori A. Loan** • **2017** • **Swiger, P. A., Patrician, P. A., Miltner, R. S. S., Raju, D., Breckenridge-Sproat, S., & Loan, L. A. (2017). The Practice Environment Scale of the Nursing Work Index: an updated review and recommendations for use. International journal of nursing studies, 74, 76-84.** • **https://www.ncbi.nlm.nih.gov/pubmed/28641123**

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	• The increased use of longitudinal and quasi-experimental designs would strengthen the evidence generated from studying the practice environment • Authors should report the internal consistency statistics for the sample they are studying and consider other techniques to evaluate instrument performance. • Authors should consider conducting a confirmatory factor analysis to test the model fit of the PES-NWI with their sample to assess the relationship between measured variables and latent variables of the instrument. • There are differences between the rating of the nurse practice environment by managers and direct care nursing staff. Authors recommend using only direct care staff nurses. • At the hospital level, this instrument demonstrated favorable content, criterion, and construct validity (CFI=0.95, and RMSA=0.057), as well as strong internal consistency (Cronbach's alphas from 0.80-0.90). • Authors need to clearly specifically what instrument they are using and if a modified instrument is being used, they should support the modification with measures of reliability and validity. • Better practice environments had lower odds of administering wrong medications, pressure ulcers and falls. • The findings of this analysis further support the importance of the nurse practice environment with regard to nurse job outcomes, patient outcomes, patient satisfaction, adverse events, and nurse-rated quality of care.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	N/A
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	N/A
Provide all other grades and definitions from the recommendation grading system	N/A

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• 46 articles from 28 countries. • Primary data collection occurred in 25 of the studies. The remaining 22 studies analyzed secondary data with the earliest reporting year of collection occurring in 1999 and the latest in 2014.
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No, the new studies make consistent conclusions with those from the systematic review.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Hospital Nurses' Work Environment Characteristics and Patient Safety Outcomes: A Literature Review** • **Seung Eun Lee and Linda D. Scott** • **2018** • **Lee, S. E., & Scott, L. D. (2018). Hospital nurses' work environment characteristics and patient safety outcomes: A literature review. Western journal of nursing research, 40(1), 121-145.** • **http://journals.sagepub.com/doi/pdf/10.1177/0193945916666071?casa_token=tzF8jZNuCFkAAAAA:uJHuW7vK9s1ObRSbPRGhWEZCjAmGdAzFTRudUpZir2nuTuxp-O**

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	• In order to clarify the relationship between the nurse work environment and patient safety outcomes researchers should use a longitudinal study design with a theoretical foundation. • There should be clear operational definitions of concepts. • The measure methodologies should be selected carefully. • Studies should provide a specific definition of the work environment. • 17 studies used cross-sectional design, which cannot reveal causal relationships between nurses' work environment characteristics and patient outcomes.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	N/A
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	N/A
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• 18 studies published between 1999 and 2016 • 17 cross-sectional studies • 1 longitudinal study
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	A newer systematic review is provided; however, conclusions are consistent.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Global Use of the Practice Environment Scale of the Nursing Work Index** • **Nora E. Warshawsky and Donna Sullivan Havens** • **2011** • **Warshawsky, N. E., & Havens, D. S. (2011). Global use of the practice environment scale of the nursing work index. Nursing research, 60(1), 17.** • **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021172/**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	• Authors recommend a reduction in scale length and consistent scoring methods. • Authors recommend the use of the measure in longitudinal and intervention research designs. • Most of the publications focused on associations between PES-NWI scores and nurse outcomes with reports consistently being associated with measures of nurse well-being through a variety of scales: job enjoyment, satisfaction, dissatisfaction, and burnout. • Associations between the nurse practice environment and the patient outcomes varied because of level of analysis, sample sizes, and measurement issues. • Careful attention should be given to unit of analysis, especially testing the performance of subscales and the composite scale at multiple levels. • More research is need to explore the association between the nurse practice environment and patient outcomes.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	N/A
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	N/A
Provide all other grades and definitions from the recommendation grading system	N/A

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• 37 research reports • 19 articles reported use of primary data and the other 18 articles did secondary analyses of eight different datasets.
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	A newer systematic review is provided; however, conclusions are consistent.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **A Meta-Analysis of the Associations Between the Nurse Work Environment in Hospitals and Five Sets of Outcomes** • **Eileen Lake, Jordan Sanders, Kathryn Riman, Kathryn Schoenauer, Rui Duan, and Yong Chen** • **Expected 2018** • **In revision following peer review at a health services research journal.** • **URL and citation not yet available.**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	• The five sets of outcomes were nurse job outcomes, safety and quality of care ratings, health record-based patient outcomes, such as mortality and hospital-acquired infection, nurse-reported adverse patient event frequency, such as patient falls, and patient satisfaction with hospital care experience. • There were consistent associations found between the work environment and all five outcomes. • Better work environments were associated with a lower odds of negative outcomes or higher odds of positive outcomes. • Hospital managers should utilize this instrument and benchmarks provided by literature to identify and address weaknesses in work environments. • Nurse education should contain content on the nurse work environment and its relation to health and job outcomes.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	N/A

Provide all other grades and definitions from the evidence grading system	The Johns Hopkins Nursing Evidence-based Practice Rating Scale was used to evaluate the strength and quality of the evidence. Newhouse R, Dearholt S, Poe S, Pugh L, White K. The Johns Hopkins Nursing Evidence-Based Practice Rating Scale. Baltimore, MD, The Johns Hopkins Hospital: 2005.
Grade assigned to the **recommendation** with definition of the grade	N/A
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• Of 324 studies published reporting empirical data on the PES-NWI since the instrument was published in 2002, through September 2018, studies were selected that met meta-analysis inclusion criteria, which were principally having a sufficient number of studies of the same outcome variable (a minimum of 3 is required for a meta-analysis model) that reported regression coefficients and confidence intervals in non-overlapping samples. • 24 studies providing observations for one or more outcome variables met inclusion criteria. • All studies were of cross-sectional design.
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This is the first meta-analysis. The conclusions are consistent with the systematic reviews.

1a.4 OTHER SOURCE OF EVIDENCE

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

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

An *in press* study (Sloane et al. 2018) provides evidence from a longitudinal panel study that changes in the nurse work environment, as measured by the PES-NWI, yield improvements in quality of care and patient safety that closely approximate the size of effects identified cross-sectionally. This study is important because:

1) The longitudinal design increases the causal basis for a link between the measure and quality outcomes.

2) The similarity in results identified in longitudinal design and a cross-sectional design imparts greater credence to the sizable cross-sectional literature showing significant associations between this measure and multiple quality outcomes.

The authors conclude: "These results are important corroboration that improving nursing resources, including the work environment, should lead to significant improvements in patient care within hospitals."(p. 2).

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

An empirical study of primary nurse survey data collected in 2006 and 2015 on a panel of 737 hospitals in 4 large U.S. states.

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

Sloane, D. M., Smith, H. L., McHugh, M. D., & Aiken, L. H. (2018). Effect of Changes in Hospital Nursing Resources on Improvements in Patient Safety and Quality of Care: A Panel Study. *Medical Care*.

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 dissemination of the PES-NWI nationally and internationally assures that nurses' practice environments will be measured in consistent fashion across different health systems to develop evidence guiding policy and management decisions. The benefit of using the PES-NWI measure for health care organizations is that organizations provide better quality patient care through improved work environments.

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.*

As noted by Warshawsky and Havens, 2011 "Using the findings and recommendations made in this review, nurse researchers can use the PES-NWI to assess nursing practice environments and to provide meaningful comparison data".

The potential range in scores is 1.00 to 4.00. This full range has been exhibited in a recent (2006) sample of 794 hospitals in 4 states. Given an average SD of .20, the score ranges across 22 studies in the table below indicate large performance gaps.

Score Ranges (Studies Reporting Scores on a 4-Point Likert Scale, n = 22)

Measure Subscale	Score Range
Collegial Nurse-Physician Relations	2.32-3.26
Nursing Foundations for Quality Care	2.20-3.35
Nurse Manager Ability, Leadership, and Support	2.08-3.42
Nurse Participation in Hospital Affairs	1.98-2.98

Staffing and Resource Adequacy 1.87-2.90

Composite 2.48-3.17

Warshawsky and Havens, 2011, Table 3

****

The Joint Commission pilot hospital PES-NWI measure rates:

	Median	Min	Max
12a) Mean score on a composite of all subscale scores	2.85	2.57	3.14
12b) Mean score on Nurse Participation in Hospital Affairs	2.74	2.33	3.09
12c) Mean score on Nursing Foundations for Quality of Care	2.96	2.67	3.28
12d) Mean score on Nurse Manager Ability, Leadership, and Support of Nurses	2.9	2.42	3.19
12e) Mean score on Staffing and Resource Adequacy	2.66	2.3	3.05
12f) Mean score on Collegial Nurse-Physician Relations	2.97	2.69	3.3

47 hospitals reported practice environment survey data, collected from August 2007 - July 2008.

In a study by the measure developer, Lake, from 794 hospitals in 4 states, the sample hospitals exhibited the full range of possible scores: 1.00 to 4.00. The average hospital-level subscale scores ranged from 2.50 to 2.84, with SDs ranging from .29 to .37.

The descriptive statistics calculated from all community hospitals in four states demonstrate lower average scores than the Joint Commission pilot hospitals as well as much greater variation across hospitals, suggesting that Joint Commission accredited hospitals have better nursing environments than all hospitals (in these 4 states and perhaps throughout the U.S.) and indicating the capacity of the PES-NWI measure to provide evidence of significant and meaningful differences in practice environmental performance across providers.

Maintenance of Endorsement (October 2018):

Descriptive statistics from the National Database of Nursing Quality Indicators

	Composite			
	mean	SD	min	max
2013	2.96	0.29	1.00	4.00
2014	2.94	0.30	1.56	4.00
2015	2.96	0.30	1.08	4.00
2016	2.99	0.30	1.56	4.00
2017	2.99	0.31	1.00	4.00
	Participation			
	mean	SD	min	max
2013	2.85	0.33	1.00	4.00
2014	2.83	0.34	1.00	4.00
2015	2.85	0.34	1.00	4.00
2016	2.88	0.33	1.44	4.00
2017	2.88	0.34	1.00	4.00
	Quality of Care			
	mean	SD	min	max
2013	3.09	0.25	1.00	4.00

2014	3.08	0.25	1.75	4.00
2015	3.09	0.26	1.13	4.00
2016	3.09	0.26	1.00	4.00
2017	3.10	0.27	1.00	4.00
	Nurse Manager			
	mean	SD	min	max
2013	2.99	0.40	1.00	4.00
2014	2.99	0.41	1.10	4.00
2015	3.02	0.41	1.00	4.00
2016	3.06	0.39	1.00	4.00
2017	3.06	0.41	1.00	4.00
	Staffing/Resources			
	mean	SD	min	max
2013	2.78	0.42	1.00	4.00
2014	2.73	0.44	1.00	4.00
2015	2.73	0.45	1.00	4.00
2016	2.73	0.45	1.00	4.00
2017	2.74	0.46	1.00	4.00
	RN/MD Relations			
	mean	SD	min	max
2013	3.10	0.32	1.00	4.00
2014	3.10	0.32	1.33	4.00
2015	3.12	0.32	1.00	4.00
2016	3.14	0.32	1.00	4.00
2017	3.15	0.33	1.00	4.00
Statistics from the Veterans Administration				
	Participation			
	mean	SD	min	max
2012	2.53	0.2	1.93	2.95
2013	2.54	0.19	2.03	3.05
2014	2.45	0.16	2.01	3
2015	2.47	0.16	2.04	2.83
2016	2.52	0.15	2.15	2.97
Quality of Care				
	mean	SD	min	max
2012	2.86	0.15	2.19	3.19
2013	2.87	0.13	2.41	3.22
2014	2.81	0.12	2.49	3.28
2015	2.81	0.12	2.45	3.06

2016 2.85 0.12 2.59 3.38

Nurse Manager

mean SD min max

2012 2.7 0.21 2 3.19

2013 2.74 0.19 2.16 3.29

2014 2.76 0.15 2.45 3.44

2015 2.82 0.13 2.48 3.21

2016 2.87 0.15 2.48 3.42

Staffing/Resources

mean SD min max

2012 2.55 0.19 1.95 2.95

2013 2.56 0.2 1.67 3.08

2014 2.58 0.14 2.25 3.07

2015 2.67 0.14 2.25 3.01

2016 2.69 0.13 2.35 3.03

2017 2.67 0.16 2.23 3.22

2018 2.69 0.18 2.22 3.47

RN/MD Relations

mean SD min max

2012 2.94 0.15 2.16 3.24

2013 2.96 0.13 2.43 3.37

2014 2.9 0.12 2.65 3.61

2015 2.91 0.11 2.58 3.18

2016 2.96 0.12 2.63 3.44

2017 3.02 0.12 2.73 3.58

2018 3.04 0.12 2.79 3.59

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.

Warshawsky, N. E., & Havens, D. S. (2011). Global use of the practice environment scale of the nursing work index. Nursing Research, 60(1), 17. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021172/

Maintenance of Endorsement (October 2018):

Add three systematic reviews and meta-analysis since 2012.

Since 2012 there have been two new systematic reviews: Swiger et al. (2017) and Lee & Scott (2018). Here are summaries demonstrating opportunity for improvement.

Swiger, P. A., Patrician, P. A., Miltner, R. S. S., Raju, D., Breckenridge-Sproat, S., & Loan, L. A. (2017). The Practice Environment Scale of the Nursing Work Index: an updated review and recommendations for use. International journal of nursing studies, 74, 76-84. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020748917301281

The literature review aimed to provide an updated review and usage recommendations for the Practice Environment Scale of the Nursing Work Index. Researchers included 46 articles from 28 countries between

2010 and 2016 that focused on the relationships between the Practice Environment Scale of the Nursing Work Index and patient, nurse, or organizational outcomes. Most studies indicated significant findings between effects of nurse practice environments on outcomes. The instrument has remained largely unchanged since its development and frequency of usage continues to be high.

This excerpt from Swiger et al. page 79 notes a performance gap in the literature:

"2.6.1. Reported PES-NWI scores

Sixteen articles (35%) reported composite PES-NWI scores, based on the 4-point Likert scale, which ranged from 2.30 to 3.07. The lowest composite score came from a study with a relatively low sample size (n =301) investigating turnover intention of registered nurses in the Eastern Caribbean who worked on medical, surgical, medical-surgical, or obstetric units (Lansiquot et al., 2012). The highest score came from a hospital in Australia that was in the process of seeking Magnet recognition (Walker et al., 2010). In studies where a sample was identified as having been collected from nurses working in Magnet facilities, the reported composite score ranged from 2.92 to 3.00 (Kutney-Lee et al., 2015; Ma and Park, 2015). Collective subscale and composite score ranges from 3 studies reporting scores from Magnet, emerging or aspiring Magnet, and non-Magnet facilities can be found in Table 1; the Staffing and Resource adequacy remains the lowest subscale for all three groups, confirming the finding from the Warshawsky and Havens

(2011) review.

This Table 1 from Swiger et al. presents score ranges from three articles demonstrating lower scores in non-Magnet hospitals, middling scores in Emerging Magnet Hospitals, and higher scores in Magnet Hospitals

Table 1 Reported Score Ranges (n = 3 articles).

PES-NWI Measure Subscale	Reported Mean Score Range (SD) Non-Magnet Scores	Emerging/Aspiring Magnet Scores	Magnet Hospital Scores
1 Nurse Participation in Hospital Affairs	2.34 - 2.87	2.49 - 3.06	2.76 - 3.01
2 Nursing Foundations for Quality of Care	2.82 - 3.11	2.98 - 3.19	3.09 - 3.20
3 Nurse Manager Ability,Leadership,& Support of Nurses	2.41 - 3.00	2.48 - 3.17	2.72 - 3.07
4 Staffing and Resource Adequacy	2.07 - 2.62	2.31 - 2.88	2.65 - 2.88
5 Collegial Nurse-Physician Relations	2.78 - 2.99	2.85 - 3.06	2.99 - 3.07
Composite	2.51 - 2.92	2.62– 3.07	2.92 - 3.00

Lee, S. E., & Scott, L. D. (2018). Hospital nurses' work environment characteristics and patient safety outcomes: A literature review. Western journal of nursing research, 40(1), 121-145. Retrieved from http://journals.sagepub.com/doi/full/10.1177/0193945916666071

The literature review conducted by Lee and Scott evaluated associations between hospital nurses' work environment characteristics and patient safety outcomes. Researchers searched five databases and reviewed 18 studies published in English between 1999 and 2016. Most studies did not include a definition for work environment, and patient outcomes were measured using different variables and instruments. The relationship between nurses' work environment characteristics and patient safety outcomes were inconsistent between studies. These 18 studies, some of which overlap with the earlier Warshawsky & Havens (2011) study, demonstrate a performance gap which was linked to patient safety outcomes

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.

Disparities not applicable to this measure.

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

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

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

https://www.jointcommission.org/assets/1/6/NSC%20Manual.pdf

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 **Attachment:** PES.NWI_Practice_Environment_Scale_of_the_Nursing_Work_Index_final_12-29-05-636511896139866185-636682914754440177.pdf

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

Clinician

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

No

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

Maintenance of Endorsement (Oct 2018): There have been no changes.

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).

Continuous Variable Statement: For surveys completed by Registered Nurses (RN):

12a) Mean score on a composite of all subscale scores

12b) Mean score on Nurse Participation in Hospital Affairs (survey item numbers 5, 6, 11, 15, 17, 21, 23, 27, 28)

12c) Mean score on Nursing Foundations for Quality of Care (survey item numbers 4, 14, 18, 19, 22, 25, 26, 29, 30, 31)

12d) Mean score on Nurse Manager Ability, Leadership, and Support of Nurses (survey item numbers 3, 7, 10, 13, 20)

12e) Mean score on Staffing and Resource Adequacy (survey item numbers 1, 8, 9, 12)

12f) Mean score on Collegial Nurse-Physician Relations (survey item numbers 2, 16, 24)

12g) Three category variable indicating favorable, mixed, or unfavorable practice environments: favorable = four or more subscale means exceed 2.5; mixed = two or three subscale means exceed 2.5; unfavorable = zero or one subscales exceed 2.5.

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 Populations:

•Registered Nurses with direct patient care responsibilities for 50% or greater of their shift

•All hospital units

•Full time, part time, and flex / pool RNs employed by the hospital

Excluded Populations

•New hires of less than 3 months

•Agency, traveler or contract nurses

•Nurses in management or supervisory roles with direct patient care responsibilities less than 50% of their shift, whose primary responsibility is administrative in nature

Data Elements by Subscale (with survey question/item number)

Nurse Participation in Hospital Affairs

PES-NWI Career Development (5)

PES-NWI Participation in Policy Decisions (6)

PES-NWI Chief Nursing Officer Visibility (11)

PES-NWI Chief Nursing Officer Authority (15)

PES-NWI Advancement Opportunities (17)

PES-NWI Administration Listens and Responds (21)

PES-NWI Staff Nurses Hospital Governance (23)

PES-NWI Nursing Committees (27)

PES-NWI Nursing Administrators Consult (28)

Nursing Foundations for Quality of Care

PES-NWI Continuing Education (4)

PES-NWI High Nursing Care Standards (14)

PES-NWI Philosophy of Nursing (18)

PES-NWI Nurses Are Competent (19)

PES-NWI Quality Assurance Program (22)

PES-NWI Preceptor Program (25)

PES-NWI Nursing Care Model (26)

PES-NWI Patient Care Plans (29)

PES-NWI Continuity of Patient Assignments (30)

PES-NWI Nursing Diagnosis (31)

Nurse Manager Ability, Leadership, and Support of Nurses

PES-NWI Supportive Supervisory Staff (3)

PES-NWI Supervisors Learning Experiences (7)

PES-NWI Nurse Manager and Leader (10)

PES-NWI Recognition (13)

PES-NWI Nurse Manager Backs up Staff (20)

Staffing and Resource Adequacy

PES-NWI Adequate Support Services (1)

PES-NWI Time to Discuss Patient Problems (8)

PES-NWI Enough Nurses for Quality Care (9)

PES-NWI Enough Staffing (12)

Collegial Nurse-Physician Relations

PES-NWI Nurse and Physician Relationships (2)

PES-NWI Nurse and Physician Teamwork (16)

PES-NWI Collaboration (24)

Composite Score

Mean of subscale scores

Three Category Variable

Favorable = four or more subscale means exceed 2.5

Mixed = two or three subscale means exceed 2.5

Unfavorable = zero or one subscales exceed 2.5

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

Staff RNs

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).

Not applicable

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

Not applicable

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.)*

Not applicable

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.)*

12a) Mean score on a composite of all subscale scores

12b) Mean score on Nurse Participation in Hospital Affairs (survey item numbers 5, 6, 11, 15, 17, 21, 23, 27, 28)

12c) Mean score on Nursing Foundations for Quality of Care (survey item numbers 4, 14, 18, 19, 22, 25, 26, 29, 30, 31)

12d) Mean score on Nurse Manager Ability, Leadership, and Support of Nurses (survey item numbers 3, 7, 10, 13, 20)

12e) Mean score on Staffing and Resource Adequacy (survey item numbers 1, 8, 9, 12)

12f) Mean score on Collegial Nurse-Physician Relations (survey item numbers 2, 16, 24)

12g) Three category variable indicating favorable, mixed, or unfavorable practice environments: favorable = four or more subscale means exceed 2.5; mixed = two or three subscale means exceed 2.5; unfavorable = zero or one subscales exceed 2.5.

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Continuous variable, e.g. average

If other:

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

Better quality = Higher score

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

1. Start processing.
2. Check Survey Date

a. If the Survey Date is missing or invalid the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.

b. If Survey Date is valid, continue and proceed to initialization.

3. Initialization. Initialize NurseParticipationScore to 0; NursingFoundationScore to 0; NurseMgrAbilityScore to 0; StaffingScore to 0; RelationsScore to 0; TotalScore to 0; ExceedCounter to 0. Continue and proceed to PES-NWI Career Development.

4. Check PES-NWI Career Development

a. If the PES-NWI Career Development is missing or zero, the case will proceed to PES-NWI Participation in Policy Decisions.

b. If the PES-NWI Career Development equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Career Development to the NurseParticipationScore and proceed to PES-NWI Participation in Policy Decisions.

5. Check PES-NWI Participation in Policy Decisions

a. If the PES-NWI-Participation in Policy Decisions is missing or zero, the case will proceed to PES-NWI Chief Nursing Officer Visibility.

b. If the PES-NWI Participation in Policy Decisions equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Participation in Policy Decisions to the NurseParticipationScore and proceed to PES-NWI Chief Nursing Officer Visibility.

6. Check PES-NWI Chief Nursing Officer Visibility

a. If the PES-NWI- Chief Nursing Officer Visibility is missing or zero, the case will proceed to PES-NWI Chief Nursing Officer Authority.

b. If the PES-NWI Chief Nursing Officer Visibility equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Chief Nursing Officer Visibility to the NurseParticipationScore and proceed to PES-NWI Chief Nursing Officer Authority.

7. Check PES-NWI Chief Nursing Officer Authority

a. If the PES-NWI- Chief Nursing Officer Authority is missing or zero, the case will proceed to PES-NWI Advancement Opportunities.

b. If the PES-NWI Chief Nursing Officer Authority equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Chief Nursing Officer Authority to the NurseParticipationScore and proceed to PES-NWI Advancement Opportunities.

8. Check PES-NWI Advancement Opportunities

a. If the PES-NWI- Advancement Opportunities is missing or zero, the case will proceed to PES-NWI Administration Listens and Responds.

b. If the PES-NWI Advancement Opportunities equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Advancement Opportunities to the NurseParticipationScore and proceed to PES-NWI Administration Listens and Responds.

9. Check PES-NWI Administration Listens and Responds

a. If the PES-NWI Administration Listens and Responds is missing or zero, the case will proceed to PES-NWI Staff Nurses Hospital Governance.

b. If the PES-NWI Administration Listens and Responds equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Administration Listens and Responds to the NurseParticipationScore and proceed to PES-NWI Staff Nurses Hospital Governance.

10. Check PES-NWI Staff Nurses Hospital Governance

a. If the PES-NWI- Staff Nurses Hospital Governance is missing or zero, the case will proceed to PES-NWI Nursing Committees.

b. If the PES-NWI Staff Nurses Hospital Governance equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Staff Nurses Hospital Governance to the NurseParticipationScore and proceed to PES-NWI Nursing Committees.

11. Check PES-NWI Nursing Committees

a. If the PES-NWI Nursing Committees is missing or zero, the case will proceed to PES-NWI Nursing Administrators Consult.

b. If the PES-NWI Nursing Committees equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nursing Committees to the NurseParticipationScore and proceed to PES-NWI Nursing Administrators Consult.

12. Check PES-NWI Nursing Administrators Consult

a. If the PES-NWI Nursing Administrators Consult is missing or zero, the case will proceed to calculate mean score on Nurse-Participation in Hospital Affairs.

b. If the PES-NWI Nursing Administrators Consult equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nursing Administrators Consult to the NurseParticipationScore and proceed to calculate mean score on Nurse-Participation in Hospital Affairs.

13. Calculate Mean Score on Nurse-Participation in Hospital Affairs. Mean Score of Nurse-Participation in Hospital Affairs equals mean of NurseParticipationScore. Assign the calculated mean score to NSC-12b. Continue and proceed to PES-NWI Continuing Education.

14. Check PES-NWI Continuing Education

a. If the PES-NWI Continuing Education is missing or zero, the case will proceed to PES-NWI High Nursing Care Standards.

b. If the PES-NWI Continuing Education equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Continuing Education to the NurseFoundationScore and proceed to PES-NWI High Nursing Care Standards.

15. Check PES-NWI High Nursing Care Standards

a. If the PES-NWI High Nursing Care Standards is missing or zero, the case will proceed to PES-NWI Philosophy of Nursing.

b. If the PES-NWI High Nursing Care Standards equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI High Nursing Care Standards to the NurseFoundationScore and proceed to PES-NWI Philosophy of Nursing.

16. Check PES-NWI Philosophy of Nursing

a. If the PES-NWI Philosophy of Nursing is missing or zero, the case will proceed to PES-NWI Nurses Are Competent.

b. If the PES-NWI Philosophy of Nursing equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Philosophy of Nursing to the NurseFoundationScore and proceed to PES-NWI Nurses Are Competent.

17. Check PES-NWI Nurses Are Competent

a. If the PES-NWI Nurses Are Competent is missing or zero, the case will proceed to PES-NWI Quality Assurance Program.

b. If the PES-NWI Nurses Are Competent equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nurses Are Competent to the NurseFoundationScore and proceed to PES-NWI Quality Assurance Program.

18. Check PES-NWI Quality Assurance Program

a. If the PES-NWI Quality Assurance Program is missing or zero, the case will proceed to PES-NWI Preceptor Program.

b. If the PES-NWI Quality Assurance Program equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Quality Assurance Program to the NurseFoundationScore and proceed to PES-NWI Preceptor Program.

19. Check PES-NWI Preceptor Program

a. If the PES-NWI Preceptor Program is missing or zero, the case will proceed to PES-NWI Nursing Care Model.

b. If the PES-NWI Preceptor Program equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Preceptor Program to the NurseFoundationScore and proceed to PES-NWI Nursing Care Model.

20. Check PES-NWI Nursing Care Model

a. If the PES-NWI Nursing Care Model is missing or zero, the case will proceed to PES-NWI Patient Care Plans.

b. If the PES-NWI Nursing Care Model equals 1, 2, 3, or 4, add the allowable value scored for Nursing Care Model to the NurseFoundationScore and proceed to PES-NWI Patient Care Plans.

21. Check PES-NWI Patient Care Plans

a. If the PES-NWI Patient Care Plans is missing or zero, the case will proceed to PES-NWI Continuity of Patient Assignments.

b. If the PES-NWI Patient Care Plans equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Patient Care Plans to the NurseFoundationScore and proceed to PES-NWI Continuity of Patient Assignments

22. Check PES-NWI Continuity of Patient Assignments

a. If the PES-NWI Continuity of Patient Assignments is missing or zero, the case will proceed to PES-NWI Nursing Diagnosis.

b. If the PES-NWI Continuity of Patient Assignments equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Continuity of Patient Assignments to the NurseFoundationScore and proceed to PES-NWI Nursing Diagnosis.

23. Check PES-NWI Nursing Diagnosis

a. If the PES-NWI Nursing Diagnosis is missing or zero, the case will proceed to calculate mean score on Nursing Foundations for Quality of Care.

b. If the PES-NWI Nursing Diagnosis equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nursing Diagnosis to theNurseFoundationScore and proceed to calculate mean score on Nursing Foundations for Quality of Care.

24. Calculate Mean Score on Nursing Foundations for Quality of Care. Mean Score of Nursing Foundations for Quality of Care equals mean of NurseFoundationScore. Assign the calculated mean score to NSC-12c. Continue and proceed to PES-NWI Supportive Supervisory Staff.

25. Check PES-NWI Supportive Supervisory Staff

a. If the PES-NWI Supportive Supervisory Staff is missing or zero, the case will proceed to PES-NWI Supervisors Learning Experience.

b. If the PES-NWI Supportive Supervisory Staff equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Supportive Supervisory Staff to the NurseMgrAbilityScore and proceed to PES-NWI Supervisors Learning Experience.

26. Check PES-NWI Supervisors Learning Experience

a. If the PES-NWI Supervisors Learning Experience is missing or zero, the case will proceed to PES-NWI Nurse Manager and Leader.

b. If the PES-NWI Supervisors Learning Experience equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Supervisors Learning Experience to the NurseMgrAbilityScore and proceed to PES-NWI Nurse Manager and Leader.

27. Check PES-NWI Nurse Manager and Leader

a. If the PES-NWI Nurse Manager and Leader is missing or zero, the case will proceed to PES-NWI Recognition.

b. If the PES-NWI Nurse Manager and Leader equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nurse Manager and Leader to the NurseMgrAbilityScore and proceed to PES-NWI Recognition.

28. Check PES-NWI Recognition

a. If the PES-NWI Recognition is missing or zero, the case will proceed to PES-NWI Nurse Manager Backs up Staff

b. If the PES-NWI Recognition equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Recognition to the NurseMgrAbilityScore and proceed to PES-NWI Nurse Manager Backs up Staff.

29. Check PES-NWI Nurse Manager Backs up Staff

a. If the PES-NWI Nurse Manager Backs up Staff is missing or zero, the case will proceed to calculate mean score on Nurse Manager Ability, Leadership, and Support of Nurses.

b. If the PES-NWI Nurse Manager Backs up Staff equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nurse Manager Backs up Staff to the NurseMgrAbilityScore and proceed to calculate mean score on Nurse Manager Ability, Leadership, and Support of Nurses.

Calculate Mean Score on Nurse Manager Ability, Leadership, and Support of Nurses. Mean Score of Nurse Manager Ability, Leadership, and Support of Nurses equals mean of NurseMgrAbilityScore. Assign the calculated mean score to NSC-12d. Continue and proceed to PES-NWI Adequate Support Services.

30. Check PES-NWI Adequate Support Services

a. If the PES-NWI Adequate Support Services is missing or zero, the case will proceed to PES-NWI Time to Discuss Patient Problems.

b. If the PES-NWI Adequate Support Services equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Adequate Support Services to the StaffingScore and proceed to PES-NWI Time to Discuss Patient Problems.

31. Check PES-NWI Time to Discuss Patient Problems

a. If the PES-NWI Time to Discuss Patient Problems is missing or zero, the case will proceed to PES-NWI Enough Nurses for Quality Care.

b. If the PES-NWI Time to Discuss Patient Problems equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Time to Discuss Patient Problems to the StaffingScore and proceed to PES-NWI Enough Nurses for Quality Care.

32. Check PES-NWI Enough Nurses for Quality Care

a. If the PES-NWI Enough Nurses for Quality Care is missing or zero, the case will proceed to PES-NWI Enough Staffing.

b. If the PES-NWI Enough Nurses for Quality Care equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Enough Nurses for Quality Care to the StaffingScore and proceed to PES-NWI Enough Staffing.

33. Check PES-NWI Enough Staffing

a. If the PES-NWI Enough Staffing is missing or zero, the case will proceed to calculate mean score on Staffing and Resource Adequacy.

b. If the PES-NWI Enough Staffing equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Enough Staffing to the StaffingScore and proceed to calculate mean score on Staffing and Resource Adequacy.

34. Calculate Mean Score on Staffing and Resource Adequacy. Mean Score of Staffing and Resource Adequacy equals mean of StaffingScore. Assign the calculated mean score to NSC-12e. Continue and proceed to PES-NWI Nurse and Physician Relationships.

35. Check PES-NWI Nurse and Physician Relationships

a. If the PES-NWI Nurse and Physician Relationships is missing or zero, the case will proceed to PES-NWI Nurse and Physician Teamwork.

b. If the PES-NWI Nurse and Physician Relationships equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nurse and Physician Relationships to the RelationsScore and proceed to PES-NWI Nurse and Physician Teamwork.

36. Check PES-NWI Nurse and Physician Teamwork

a. If the PES-NWI Nurse and Physician Teamwork is missing or zero, the case will proceed to PES-NWI Collaboration.

b. If the PES-NWI Nurse and Physician Teamwork equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Nurse and Physician Teamwork to the RelationsScore and proceed to PES-NWI Collaboration.

37. Check PES-NWI Collaboration

a. If the PES-NWI Collaboration is missing or zero, the case will proceed to calculate mean score on Collegial Nurse-Physician Relations.

b. If the PES-NWI Collaboration equals 1, 2, 3, or 4, add the allowable value scored for PES-NWI Collaboration to the RelationsScore and proceed to calculate mean score on Collegial Nurse-Physician Relations.

38. Calculate Mean Score on Collegial Nurse-Physician Relations. Mean Score of Collegial Nurse-Physician Relations equals mean of RelationsScore. Assign the calculated mean score to NSC-12f. Continue and proceed to calculate the Total Score on composite of all subscale scores.

39. Calculate Total Score on a composite of all subscale scores. Total Score of a composite of all subscale scores equals the sum of NurseParticipationScore, NursingFoundationScore, NurseMgrAbilityScore, StaffingScore, and RelationsScore. Continue and proceed to calculate Mean Score on a composite of all subscale scores.

40. Calculate Mean Score on a composite of all subscale scores. Mean Score of a composite of all subscale scores equals the mean of Total Score on a composite of all subscale scores. Assign the calculated mean score to NSC-12a. Continue and proceed to Mean Score on NurseParticipationScore.

41. Check Mean Score on NurseParticipationScore

a. If the score of Mean Score on NurseParticipationScore is less than or equal to 2.5, the case will proceed to Mean Score on NursingFoundationScore.

b. If the score of Mean Score on NurseParticipationScore is greater than 2.5, add 1 to ExceedCounter and proceed to Mean Score on NursingFoundationScore.

42. Check Mean Score on NursingFoundationScore

a. If the score of Mean Score on NursingFoundationScore is less than or equal to 2.5, the case will proceed to Mean Score on NurseMgrAbilityScore.

b. If the score of Mean Score on NursingFoundationScore is greater than 2.5, add 1 to ExceedCounter and proceed to Mean Score on NurseMgrAbilityScore.

43. Check Mean Score on NurseMgrAbilityScore

a. If the score of Mean Score on NurseMgrAbilityScore is less than or equal to 2.5, the case will proceed to Mean Score on StaffingScore.

b. If the score of Mean Score on NurseMgrAbilityScore is greater than 2.5, add 1 to ExceedCounter and proceed to Mean Score on StaffingScore.

44. Check Mean Score on StaffingScore

a. If the score of Mean Score on StaffingScore is less than or equal to 2.5, the case will proceed to Mean Score on RelationsScore.

b. If the score of Mean Score on StaffingScore is greater than 2.5, add 1 to ExceedCounter and proceed to Mean Score on RelationsScore.

45. Check Mean Score on RelationsScore

a. If the score of Mean Score on RelationsScore is less than or equal to 2.5, the case will proceed to ExceedCounter.

b. If the score of Mean Score on RelationsScore is greater than 2.5, add 1 to ExceedCounter and proceed to ExceedCounter.

46. Check ExceedCounter

a. If ExceedCounter is greater than or equal to 4, the case will proceed to a Measure Category Assignment of "Favorable". Stop processing.

b. If ExceedCounter is greater than or equal to 2 and less than 4, the case will proceed to a Measure Category Assignment of "Mixed". Stop processing.

c. If ExceedCounter is greater than or equal to 0 and less than 2, the case will proceed to a Measure Category Assignment of "Unfavorable". Stop processing.

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.

For public reporting, the specific sampling approach is a random sample of 50 direct care staff registered nurses. With an anticipated response rate of 60%, the publicly reported measure would be based on 30 or more responses. The minimum of 30 is based on The Joint Commission´s established minimum for comparative results to be calculated to represent the hospital. Satisfactory estimates of PES hospital scores have been obtained with fewer than 30 responses (Lake & Friese, 2006). Nevertheless, a larger sample improves the precision of the results. While a random sample may be used at the hospital-level, it is recommended that hospitals survey all eligible nurses to allow all nurses the opportunity to complete the practice environment survey instrument.

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.

According to Lake and Friese (2006) the minimum number of completed surveys per hospital for satisfactory estimates is 15, therefore considering a typical response rate of 60%, a random sample of at least 25 nurses needs to be surveyed annually. For purposes of public reporting the measure a minimum of 30 completed surveys is desired, therefore hospitals that choose to sample should sample a minimum of 50 nurses annually. While a random sample may be used at the hospital-level, it is recommended that hospitals survey all eligible nurses to allow all nurses the opportunity to complete the practice environment survey instrument.

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.

Practice Environment Scale-Nursing Work Index (PES-NWI) 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 at measure-specific web page URL identified in S.1

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

Clinician : Group/Practice, Facility

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

Inpatient/Hospital

If other:

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

2. Validity – See attached Measure Testing Submission Form

0206_MeasureTesting_CompositeMSF1.0_Data-636682914759440716.doc,0206.nqf_testing_attachment_7.1_july.31.pdf

2.1 <u>For maintenance of endorsement</u>

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 <u>For maintenance of endorsement</u>

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 <u>For maintenance of endorsement</u>

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.

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 3450

Measure Title: Practice Environment Scale - Nursing Work Index (PES-NWI) (composite and five subscales)

Date of Submission: 7/31/2018

Type of Measure:

☐ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource

☐ Process *(including Appropriate Use)*	☐ Efficiency
☒ Structure	

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins).

  Contact NQF staff if more pages are needed.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument- based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: nurse survey	☒ other: nurse survey

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*).

Nurse survey data from research projects and the National Database of Nursing Quality Indicators were used to derive and confirm the instrument subscales and composite and to provide ongoing psychometric performance.

1.3. What are the dates of the data used in testing? 1985 to 2018

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.20*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

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

Most measured entities were acute care hospitals. Some measured entities were home care agencies. Per Warhsawsky & Havens (2010), 37 samples of 4 to 4,783 units over the years 1998 to 2010, and per Swiger, et. al 2017, 46 samples of 2 units to 5322 units and 519 hospitals over the years 2010 to 2016. In addition, per Lake et al. 2018, 212 separate research articles were published through March 2016 that included empirical data on the PES-NWI; some of these articles were included in the two systematic reviews noted previously. From April 2016 through June 2018, 35 separate research articles were published that included empirical data on the PES-NWI from 7 to 489 hospitals. These hospital samples included representative samples of hospitals from multiple U.S. states, including hospitals of all sizes, ownership, and teaching status.

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*)

There are no patient data. Here we report about data from nurses. Per Warhsawsky & Havens (2010), 37 samples of 31 to 72,889 nurses over the years 1998 to 2010, and per Swiger, et. al 2017, 46 samples of 133 to 33,845 over the years 2010 to 2016. In addition, per Lake et al. 2018, 212 separate research articles were published through March 2016 that included empirical data on the PES-NWI; some of these articles were included in the two systematic reviews noted previously. From April 2016 through June 2018, 37 separate research articles were published that included empirical data on the PES-NWI. In these 35 articles, data from samples of from 87 to 33,000 nurses were reported. The nurse characteristics in many samples resembled nurse characteristics for age, sex, and educational level as described in national nurse surveys.

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 are no differences for different aspects of testing.

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.

There is no basis for adjusting for social factors of nurses, such as educational level. There is no contextual reason to think that social factors of nurses would impact their answers or impact being able to compare facilities fairly.

2a2. RELIABILITY TESTING

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

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

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

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

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability testing of critical data elements was conducted by computing Cronbach's alpha, which measures internal consistency of the items in a scale. This method was computed and reported in all studies noted above in the Warshawsky & Havens (2010) and Swiger et. al, 2017 papers.

Reliability testing of performance measure score was conducted by assessing inter-rater reliability, which focuses on whether nurses give consistent responses within a hospital or nursing unit, as compared to across hospitals or nursing units in a sample. Performance measure score reliability is assessed using the intraclass correlation (ICC) (1,*k*), which is a function of the number of nurse respondents per hospital and the intraclass correlation coefficient from a one-way analysis of variance of the subscales and composite across hospitals or nursing units. In order to assure reliability, the ICC (1,*k*) should exceed .60 (Glick, 1985).

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, 10(3), 601-616.

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

Reliability testing of Critical data elements: Cronbach's Alpha Statistics

Of the 46 articles reviewed in Swiger et al (2017) published from 2010 to 2016, 37 reported Cronbach's alphas; coefficients ranged from .71 – .96, with the exception of one .67, and one .53 in a small sample size. These results support the coherence of the different subscales and the composite. Additional internal consistency reliability data are displayed in Table 2b1.3D, from the 35 newest articles. This table is presented at the end of the document due to its length of several pages.

Distribution of Reliability Statistics from a Signal-To-Noise Analysis: Statistics on Organizational Reliability:

Table 2a2.3A.

Analysis for 2018 NQF measure maintenance using 2015 National Database of Nursing Quality Indicators nurse survey data

Measure	ICC (1,k)
Subscale	
Collegial Nurse-Physician Relations	.936
Nursing Foundations for Quality Care	.966
Nurse Manager Ability, Leadership, and Support	.949
Nurse Participation in Hospital Affairs	.973
Staffing and Resource Adequacy	.967
Composite	.966

Note. N = 451 hospitals and from 157,481 to 157,522 staff nurses. ICC (1,k) estimated in one-way ANOVA.

Table 2a2.3B.

Compilation of entity-level reliability statistics across 14 studies published from 2002 to 2017.

References:

Reference	# organizational units (hospitals or nursing units)	# nurses	ICC (1,k) statistics reported or summarized	Page reference
Lake (2002)	16 magnet hospitals proportionate by regions of the country	1,610	.88 to .97	Pg 183
Lake et al (2006)	156 adult community hospitals in Pennsylvania	10,962	.67 to .82	Pg 4
Clarke (2007)	188 Pennsylvania general acute care hospitals	11,512	.70 to .90	Pg 303
Flynn et al (2010)	63 Medicare and Medicaid certified nursing homes in New Jersey	897	Composite: .68 Subscales range: .55 to .75	Pg 4, 9
Brooks- Carthon et al (2011)	429 hospitals across four states (Florida, Pennsylvania, New Jersey and California)	98,000	Subscales range: .73 to .90	Pg 303
McHugh et al (2012)	396 adult, non-federal acute care hospitals across four states (CA, FL, NJ, PA)	16,241	.61	Pg 3
Kelly et al (2013)	320 hospitals across four states (CA, FL, NJ, PA)	3,217	.69	Pg 484
McHugh et al (2013)	564 Magnet and non-Magnet hospitals across four states (CA, FL, NJ, PA)	100,000	.81	Pg 4
Kelly et al (2014)	303 adult care hospitals across four states (CA, FL, NJ, PA)	55,159	.71	Pg 4
McHugh et al (2014)	534 hospitals across four states (CA, FL, NJ, PA)	26,005	.85	Pg 74
Carthon et al (2015)	419 acute care hospitals across three states (CA, FL, NJ, PA)	20,605	.74 to .91	Pg 257
Ma et al (2015)	373 hospitals from 44 states	33,845	Ranged from .80 to .87	Pg 3
Lake et al (2016)	171 hospitals across four states (CA, FL, NJ, PA)	1,247	4 subscales >.60; 5th = .58	Pg 3
Swiger et al (2018)	45 acute care units in 10 Army hospitals	180	ICC (1,*k*) reported as satisfactory	Pg 134, 136

Lake, E. T. (2002). Development of the practice environment scale of the Nursing Work Index. Research in nursing & health, 25(3), 176-188.

Lake, E. T., & Friese, C. R. (2006). Variations in nursing practice environments: relation to staffing and hospital characteristics. Nursing research, 55(1), 1-9.

Clarke, S. P. (2007). Hospital work environments, nurse characteristics, and sharps injuries. American Journal of Infection Control, 35(5), 302-309.

Flynn, L., Liang, Y., Dickson, G. L., & Aiken, L. H. (2010). Effects of nursing practice environments on quality outcomes in nursing homes. Journal of the American Geriatrics Society, 58(12), 2401-2406.

Brooks-Carthon, J. M., Kutney-Lee, A., Sloane, D. M., Cimiotti, J. P., & Aiken, L. H. (2011). Quality of care and patient satisfaction in hospitals with high concentrations of black patients. Journal of Nursing Scholarship, 43(3), 301-310.

McHugh, M. D., & Stimpfel, A. W. (2012). Nurse reported quality of care: a measure of hospital quality. Research in nursing & health, 35(6), 566-575.

Kelly, D., Kutney-Lee, A., Lake, E. T., & Aiken, L. H. (2013). The critical care work environment and nurse-reported health care–associated infections. American Journal of Critical Care, 22(6), 482-488.

McHugh, M. D., Kelly, L. A., Smith, H. L., Wu, E. S., Vanak, J. M., & Aiken, L. H. (2013). Lower mortality in magnet hospitals. Medical care, 51(5), 382.

Kelly, D. M., Kutney-Lee, A., McHugh, M. D., Sloane, D. M., & Aiken, L. H. (2014). Impact of critical care nursing on 30-day mortality of mechanically ventilated older adults. Critical care medicine, 42(5), 1089.

McHugh, M. D., & Ma, C. (2014). Wage, work environment, and staffing: effects on nurse outcomes. Policy, Politics, & Nursing Practice, 15(3-4), 72-80.

Carthon, J. M. B., Lasater, K. B., Sloane, D. M., & Kutney-Lee, A. (2015). The quality of hospital work environments and missed nursing care is linked to heart failure readmissions: a cross-sectional study of US hospitals. BMJ Qual Saf, 24(4), 255-263.

Ma, C., & Park, S. H. (2015). Hospital magnet status, unit work environment, and pressure ulcers. Journal of Nursing Scholarship, 47(6), 565-573.

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?*)

The measure exhibits high internal consistency reliability as well as high performance score reliability, exemplified through satisfactory ICC(1,*k*) values in 14 samples over 16 years, plus recent 2015 national data from 157,500 nurses in 451 hospitals analyzed for NQF measure maintenance.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ Performance measure score

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality

or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The method of validity testing was by statistical association between the measure and hypothesized related constructs, to demonstrate construct, concurrent, and predictive validity.

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

The PES-NWI was developed in 2002 to measure nursing practice environments through factor analysis of 1986 survey data from staff nurses in 16 original magnet hospitals, and confirmed in 1999 data from 11,636 nurses throughout Pennsylvania (Lake, 2002). The five PES-NWI subscales can be combined into a composite

measure of the practice environment, as either a continuous variable or a three-category variable indicating favorable, mixed, or unfavorable practice environments (Lake & Friese, 2006).

Lake, E. T. (2002). Development of the practice environment scale of the Nursing Work Index. Research in nursing & health, 25(3), 176-188.

Lake, E. T., & Friese, C. R. (2006). Variations in nursing practice environments: relation to staffing and hospital characteristics. Nursing research, 55(1), 1-9.

Validity testing since the measure was developed entails evaluating hypothesized relationships by computing correlation coefficients, ANOVAs, t-tests and estimating regression coefficients.

Here we describe these associations as summarized from two systematic reviews.

Warshawsky & Havens (2011) report that the majority of the 37 studies associated the PES-NWI with organization (n = 16 studies), nurse outcomes (n = 23 studies), or patient outcomes (n = 16 studies). Studies reported nurse outcomes including, job satisfaction, intent to leave, burnout, and work engagement. Articles reported patient related outcomes including, patient satisfaction, and medication errors. Moreover, studies investigated organizational outcomes such as safety climate and morale. The results of these analyses are displayed in Warshawsky & Havens Table 4 on article pages 10 & 11.

Swiger et al. (2017) report that the majority of the 46 studies they reviewed associated the PES-NWI with organization (n = 8 studies), nurse outcomes (n = 24 studies), or patient outcomes (n = 14 studies).

Scores in Magnet and Non-Magnet Hospitals Demonstrating Discriminant Validity

We hypothesize that work environments in Magnet hospitals, recognized for achieving excellent nursing standards, will have higher scores than work environments in non-Magnet hospitals. In this table, present the score ranges by Magnet status. In Table 2b1.3A we show studies where data were collected from nurses working in Magnet hospitals and non-Magnet hospitals. We show that scores were significantly higher in the Magnet facilities, demonstrating the continued discriminant ability of the instrument.

Table 2b1.3A.

Replication of Swiger et al., 2017 (Table 1): Reported Score Ranges (n = 3 articles)

PES-NWI Measure	**Reported Mean Score Range (SD)**	
Subscale	Non-magnet scores	Magnet hospital scores
Nurse participation in hospital affairs	2.34-2.87	2.76-3.01
Nursing foundations for quality of care	2.82-3.11	3.09-3.20
Nurse manager ability, leadership, & support of nurses	2.41-3.00	2.72-3.07
Staffing and resource adequacy	2.07-2.62	2.65-2.88
Collegial nurse-physician relations	2.78-2.99	2.99-3.07
Composite	2.51-2.92	2.92-3.00

Additionally, of the 13 publications that reported PES-NWI composite scores studied by Warshawsky and Havens, the lowest score reported (2.48) was by acute care nurses working in non-Magnet hospitals in Pennsylvania (Lake, 2002). Furthermore, three studies reported positive correlations between PES-NWI scores and Magnet hospital recognition (Friese et al., 2005; Lake, 2002; Lake & Friese, 2006).

Studies noted above:

Kelly, L. A., McHugh, M. D., & Aiken, L. H. (2011). Nurse outcomes in Magnet® and non-Magnet hospitals. The Journal of nursing administration, 41(10), 428.

Kutney-Lee, A., Stimpfel, A.W., Sloane, D.M., Cimiotti, J.P., Quinn, L.W., Aiken, L.H., 2015. Changes in patient and nurse outcomes associated with magnet hospital recognition. Med. Care 53 (6), 550–557.

Ma, C., Park, S.H., 2015. Hospital magnet status, unit work environment, and pressure ulcers. J. Nurs. Scholarsh. 47 (6), 565–573

McHugh, M. D., Kelly, L. A., Smith, H. L., Wu, E. S., Vanak, J. M., & Aiken, L. H. (2013). Lower mortality in magnet hospitals. Medical care, 51(5), 382.

Walker, K., Middleton, S., Rolley, J., Duff, J., 2010. Nurses report a healthy culture: results of the Practice Environment Scale (Australia) in an Australian hospital seeking Magnet recognition. Int. J. Nurs. Pract. 16 (6), 616–623.

In Table 2b1.3B we note the hypothesized relationship with the various outcomes and report the studies linking the PES-NWI to those outcomes from the two systematic reviews. The last column shows the direction of the association (- or +) and the value of the coefficients. Evidence from the 35 studies published since the later systematic review is presented in Table 2b1.3D at the end of the document for ease of viewing.

Table 2b1.3B

Statistical evidence of associations between the PES-NWI and related constructs

Outcomes	**Hypothesized relationship with PES-NWI**	**Research study**	**Statistical test value**
Patient Record Outcomes			
30 day inpatient mortality	negative		
		Aiken et al (2008)	(-, OR = 0.91)
		Aiken et al (2011) b	(-, OR = 0.93)
		Cho et al (2014)	(-, OR = 0.52)
		Friese et al (2008)	- Mortality
		Nicely et al (2013)	(-, OR = 0.89)
		Kelly (2014)	(-, OR = 0.97)
30 day hospital readmission	negative		
		Gardner et al (2007)	— Hospitalizations
		Ma & Park (2015)	(-, OR = 0.97)
		McHugh et al (2016)	(-, OR = 0.84)
Complications	negative		
		Friese et al (2008)	-
Failure to rescue	negative		
		Aiken et al (2008)	(-, OR = 0.91)
		Aiken et al (2011) b	(-, OR = 0.93)
		Friese et al (2008)	- Failure to rescue
		Nicely et al (2013)	(-, OR = 0.90)
Discharged without breastmilk	negative		
		Lake (2016)	(-, OR= 0.92)
Percent of infants on unit discharged on breastmilk	positive		
		Hallowell et al (2016)	(+, β = 0.04); Adjusted R2 = 0.37
Nurse-reported (NR) Adverse Outcomes			
NR nosocomial infection	negative		
		Kutney-Lee et al (2009)	-

Outcomes	Hypothesized relationship with PES-NWI	Research study	Statistical test value
		Lake et al (2015)	(-, OR= 0.85)
		Spence Laschinger and Leiter (2006)	-
NR patient falls	negative		
		Cho et al (2016)	Falls with injury (-, OR = 0.68)
		Kutney-Lee, Lake, et al (2009)	- Falls with injury
		Prezerakos et al (2015)	All falls (-, OR= .02)
		Spence Laschinger & Leiter (2006)	- All falls
NR medication errors	negative		
		Cho et al (2016)	(-, OR=0.55)
		Manojlovich & DeCicco (2007)	-
		Spence Laschinger & Leiter (2006)	-
NR catheter-associated sepsis	negative		
		Manojlovich & DeCicco (2007)	-
NR pressure ulcer	negative		
		Cho et al (2016)	(-, OR = 0.61)
		Choi and Staggs (2014)	Unit acquired pressure ulcers SRA (-, OR = 0.78)
		Flynn et al (2010)	(-, β=0.37)
		Ma and Park (2015)	(-, OR= 0.73)
NR urinary tract infection	negative		
		Kelly (2013)	(-, OR= 0.80)
NR bloodstream infection	negative		
		Kelly (2013)	(-, OR=0.77)
NR pneumonia	negative		
		Kelly (2013)	(-, OR= 0.80)
NR central line infection	negative		
		Lake (2016)	(-, OR= 0.89)
Patient Satisfaction			
Patient safety climate	positive		
		Armstrong and Laschinger (2006)	+
		Armstrong et al (2009)	+
Perceived quality of care	positive		
		Gardner et al (2009)	+
Nurses communicated well	positive		
		Aiken et al (2012)	(+, OR=1.11)
		You et al (2013)	(+, OR= 1.30)
Patient rates hospital highly	positive		
		Aiken et al (2012)	(+, OR= 1.16)
		Kutney-Lee et al (2015)	(+, OR= 1.17)
		You et al (2013)	(+, OR= 1.29)
Patient satisfaction	positive		
		Boev (2012)	Patient satisfaction and Nurse Manager Ability and Support of Nurses (+, β= 0.424)
		Kutney-Lee, McHugh, et al (2009)	+ (HCAHPS)

Outcomes	Hypothesized relationship with PES-NWI	Research study	Statistical test value
		Tei-Tominaga and Sato (2016)	and NPR (+, OR= 0.144)

Table 2b1.3C below reports mean and range for percentage of patients who reported on the variables indicated and regression coefficient from a linear regression of the HCAHPS variable on the PES-NWI composite score.

Table 2b1.3C

Analysis for 2018 NQF measure maintenance Linking 2015 hospital-level data from the Hospital Consumer Assessment of Health Providers and Systems HCAHPS to the PES-NWI from the National Database of Nursing Quality Indicators (n = 390).

HCAPHS Measure	Measure definition	M	Range	β coefficient and 95% CI
Composite Measures				
Communication with nurses	Patients who reported that their nurses "Always" communicated well	79	63 – 93	9.75*** (7.65-11.86)
Responsiveness of hospital staff	Patients who reported that they "Always" received help as soon as they wanted	65	44 – 86	14.30*** (10.76-17.83)
Pain management	Patients who reported that their pain was "Always" well controlled	70	56 – 84	11.21*** (9.05-13.37)
Communication about medicines	Patients who reported that staff "Always" explained about medicines before giving it to them	64	53 – 81	11.27*** (8.88-13.67)
Discharge information	Patients who reported that YES, they were given information about what to do during their recovery at home	87	79 – 97	3.87*** (2.11-5.63)
Care Transition	Patients who "Strongly Agree" they understood their care when they left the hospital	52	33 – 69	14.60*** (11.74-17.45)
Global measures				
Overall rating of hospital	Patients who gave their hospital a rating of 9 or 10 on a scale from 0 (lowest) to 10 (highest)	70	50 - 95	18.98*** (15.06–22.90)
Willingness to recommend the hospital	Patients who reported YES, they would definitely recommend the hospital	73	44 – 98	20.73*** (16.20-25.27)

Note. N = 390 hospitals except for overall rating of hospital (n = 377 hospitals); ***p < .001

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 results demonstrate that the measure exhibits satisfactory validity across a wide range of related constructs in many international samples across 16 years as well as in national 2015 data analyzed for measure re-endorsement.

2b2. EXCLUSIONS ANALYSIS

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

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

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*)

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*)

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.

Risk adjustment is not applicable

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with 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.

Not applicable

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.

The conceptual rationale for not controlling for differences in nurse characteristics is that nurse capacity to assess aspects of the work environment does not depend on nurse age, sex, or educational level. All nurses in direct clinical care positions are ideally positioned to make these assessments.

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?

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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.**

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

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

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**: 2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*): **2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**: **2b3.9. Results of Risk Stratification Analysis**:

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*)

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 method was to provide descriptive statistics at the level of the measured entities (hospitals or nursing units) showing mean, standard deviation, and range.

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*)

Warshawsky & Havens (2011) reported PES-NWI scores on a 4-point Likert Scale across 22 studies. The theoretical range is from 1.00 to 4.00. The composite score range was reported as 2.48 to 3.17. The subscale score ranges are demonstrated in Table 3 of Warshawsky & Havens, replicated here:

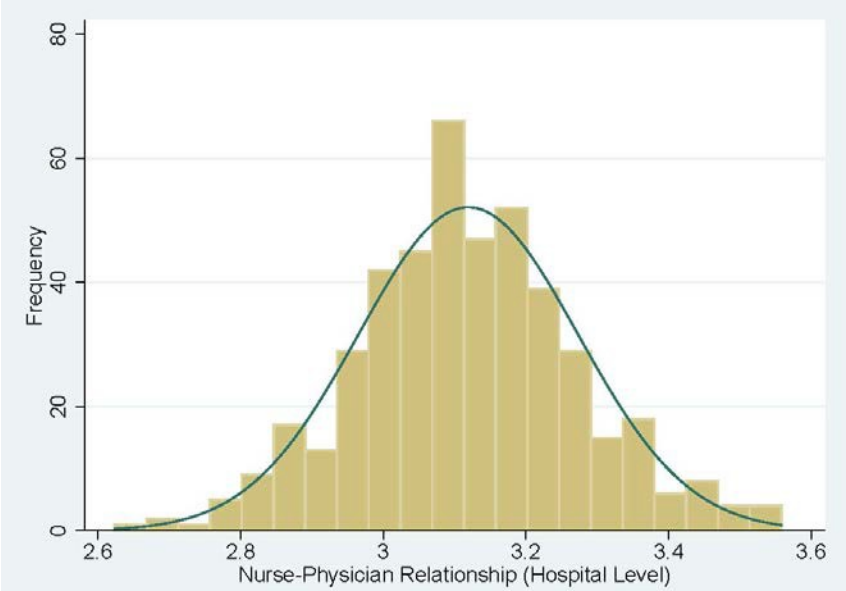
Measure	Score Range
Subscale	
Collegial Nurse-Physician Relations	2.32-3.26
Nursing Foundations for Quality Care	2.20-3.35
Nurse Manager Ability, Leadership, and Support	2.08-3.42
Nurse Participation in Hospital Affairs	1.98-2.90
Staffing and Resource Adequacy	1.87-2.90
Composite	2.48-3.17

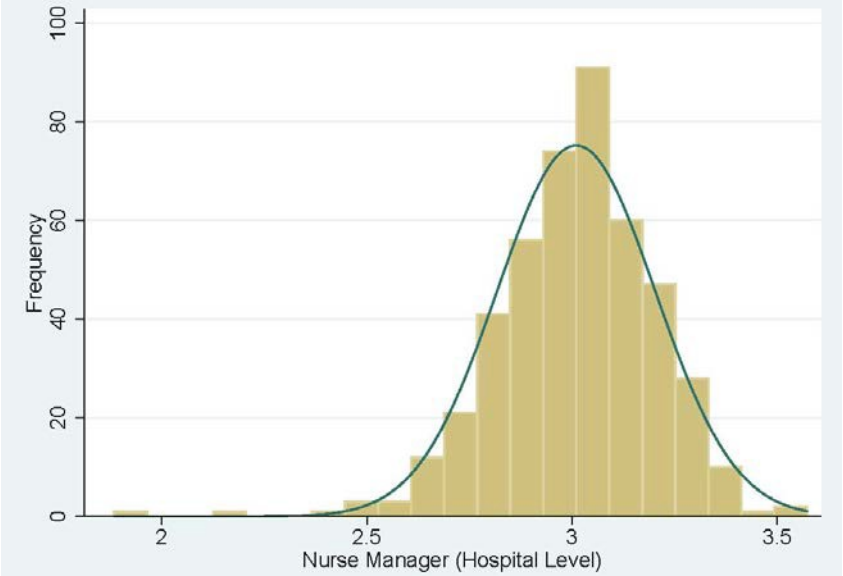
Warshawsky, N. E., & Havens, D. S. (2011). Global use of the practice environment scale of the nursing work index. Nursing research, 60(1), 17.

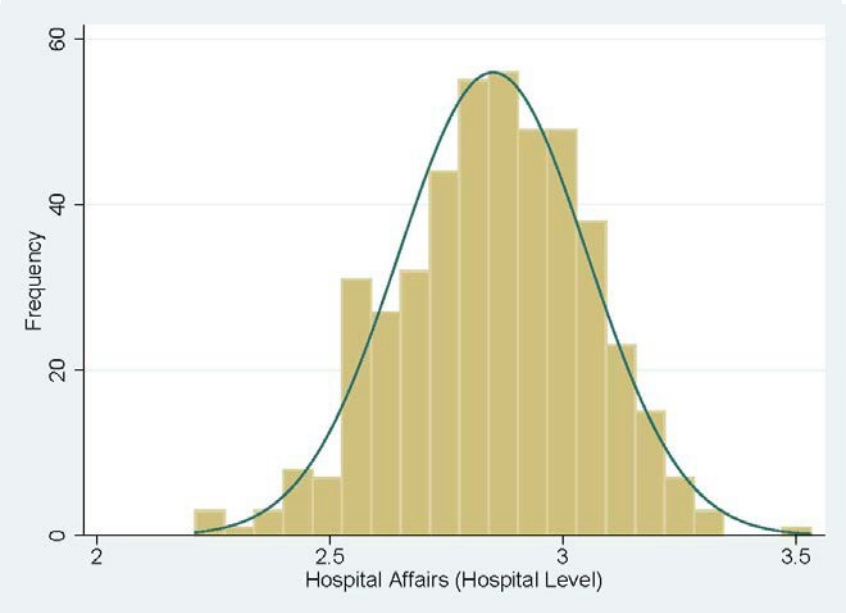
In a 2017 review of the PES-NWI measure (Swiger), sixteen articles reported composite scores ranging from 2.30 to 3.07 based on the 4-point Likert scale. Composite scores showed meaningful variation. Like in Warshawsky & Havens (2011), the Staffing and Resource Adequacy subscale remains the lowest range for hospitals.

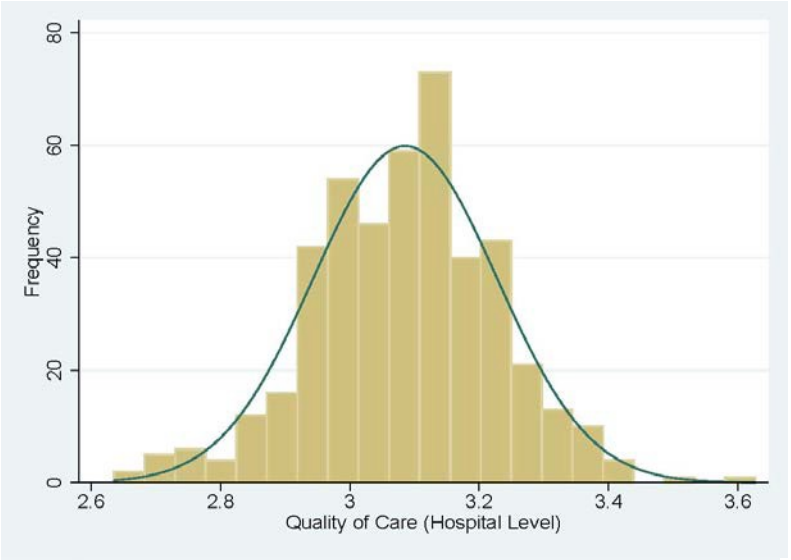
Swiger, P. A., Patrician, P. A., Miltner, R. S. S., Raju, D., Breckenridge-Sproat, S., & Loan, L. A. (2017). The Practice Environment Scale of the Nursing Work Index: an updated review and recommendations for use. International journal of nursing studies, 74, 76-84.

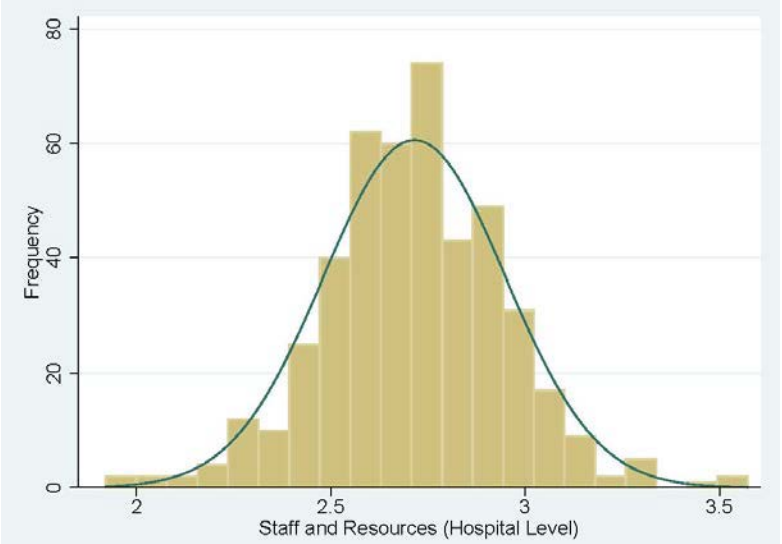
Analysis for 2018 NQF measure maintenance: Density plots displayed below for each subscale and composite measure of the PES-NWI provided from 2015 NDNQI data of 452 hospitals provide further insight to the meaningful differences in measure scores. The differences in the distributions across subscales show that they provide meaningful measures for comparison across hospitals of constructs that may be targets for institutional improvements.

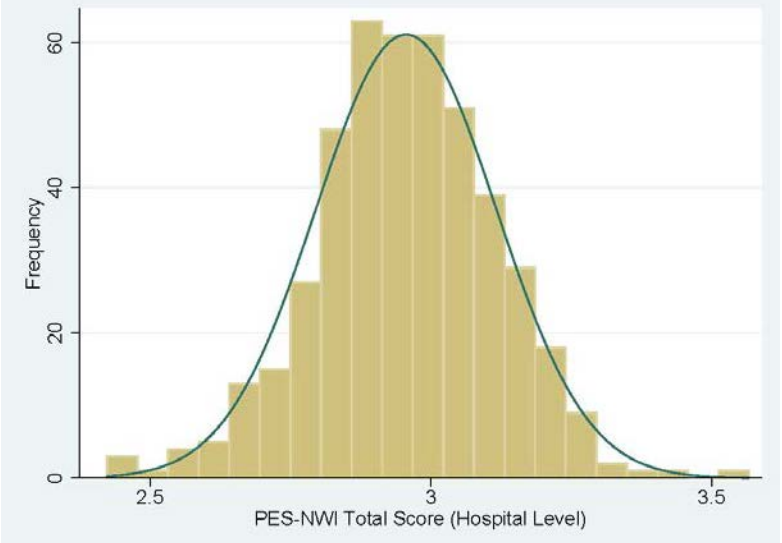

Frequency
0
20
40
60
80
2.6
2.8
3
3.2
3.4
3.6
Nurse-Physician Relationship (Hospital Level)


Frequency
0
20
40
60
80
100
2
2.5
3
3.5
Nurse Manager (Hospital Level)


Frequency
0
20
40
60
2
2.5
3
3.5
Hospital Affairs (Hospital Level)


Frequency
0
20
40
60
80
2.6
2.8
3
3.2
3.4
3.6
Quality of Care (Hospital Level)


Frequency
0
20
40
60
80
2
2.5
3
3.5
Staff and Resources (Hospital Level)


Frequency
0
20
40
60
2.5
3
3.5
PES-NWI Total Score (Hospital Level)

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?*)

There are consistent statistically significant and clinically meaningful differences in performance across measured entities.

A unique study measured changes in the PES-NWI composite score in a panel of hospitals from 1999 to 2006 (Kutney-Lee et al. 2013). This study demonstrates that work environments can change over time, which provides the basis for improving work environments in order to enhance quality of care and patient outcomes. The study also demonstrated that improvements in work environments had a strong negative association with changes in rates of job dissatisfaction, nurse burnout, and intention to leave the job. These are the relationships that have been observed in cross-sectional studies. The finding in a longitudinal design enhances the causal basis for this structural element to influence care quality and nurse and patient outcomes.

Kutney-Lee, A., Wu, E. S., Sloane, D. M., & Aiken, L. H. (2013). Changes in hospital nurse work environments and nurse job outcomes: an analysis of panel data. *International journal of nursing studies*, *50*(2), 195-201.

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.***

Not applicable

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*)

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*)

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*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Analysis for 2018 NQF measure maintenance: Missing data were calculated for the 31 items that comprise the measure.

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*)

Statistics from 2015 NDNQI nurse survey data: For each of the 31 items: At the respondent level: less than 1% of respondents have missing data. At the hospital level, about 90% of hospitals have less than 4% of their respondents with missing data.

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

Our interpretation is that missing data is minimal and appears to be at random. Therefore, performance results would be non-biased.

Table 2b1.3D

Reference	# hospitals	# nurses	Outcome measure	Reliability (Cronbach's alpha)
Yan P, Yang Y, Zhang L, et al. Correlation analysis between work-related musculoskeletal disorders and the nursing practice environment, quality of life, and social support in the nursing professionals. Medicine. 2018;97(9):e0026.	12 hospitals	2170 nurses	Work related musculoskeletal disorders	0.91 for composite 0.67-0.79 for the subscale Retest reliability was 0.84 Content validity was 0.94
Wu Y, Zheng J, Liu K, et al. The associations of occupational hazards and injuries with work environments and overtime for nurses in China. Res Nurs Health. 2018.	111 medical/surgical units in 23 hospitals	1517 nurses	Occupational hazards and injuries	0.96 for composite 0.79-0.93 for the subscales
Wan Q, Zhou W, Li Z, Shang S, Yu F. Work engagement and its predictors in registered nurses: A cross-sectional design. Nurs Health Sci. 2018.	10-15 units in 3 specialized hospitals	1065 registered nurses	Work engagement	0.89 for composite 0.60-0.75 for the subscales
Swiger PA, Loan LA, Raju D, Breckenridge- Sproat ST, Miltner RS, Patrician PA. Relationships between Army nursing practice environments and patient outcomes. Res Nurs Health. 2018;41(2):131- 144	45 units in 10 Army hospitals	1,710 of all nurse types	Patient outcomes (falls, medication errors, etc.)	0.94-0.95 for the composite 0.79-0.91 for the subscale
Smith JG, Morin KH, Lake ET. Association of the nurse work environment with nurse incivility in hospitals. J Nurs Manag. 2018;26(2):219-226.	5 acute care hospitals	233 staff nurses	Work incivility	0.94 for the composite 0.83-0.86 for the subscales
Newhouse R, Byon HD, Storkman Wolf E, Johantgen M. Multisite Studies Demonstrate Positive Relationship Between Practice Environments and Smoking Cessation Counseling Evidence-Based Practices. Worldviews Evid Based Nurs. 2018;15(3):217-224.	45 hospitals	844 registe red nurses	Nurse smoking cessation counseling practices	no
Nelson-Brantley HV, Park SH, Bergquist- Beringer S. Characteristics of the Nursing Practice Environment Associated With Lower Unit-Level RN Turnover. The Journal of nursing administration. 2018;48(1):31-37.	1002 adult care units in 162 NDNQI hospitals	Does not report	RN turnover	0.82 for the composite α≥ 0.80 for the subscales, with the exception of the interprofessional relations subscale (α= 0.71)

Reference	# hospitals	# nurses	Outcome measure	Reliability (Cronbach's alpha)
Moreno-Casbas MT, Alonso-Poncelas E, Gomez-Garcia T, Martinez-Madrid MJ, Escobar-Aguilar G. Perception of the quality of care, work environment and sleep characteristics of nurses working in the National Health System. Enferm Clin. 2018.	7 hospitals	635 registered nurses	Measure relationship between ward and work shift with nurses' perception their work environment, and sleep quality	no
Hiler CA, Hickman RL, Jr., Reimer AP, Wilson K. Predictors of Moral Distress in a US Sample of Critical Care Nurses. American journal of critical care : an official publication, American Association of Critical- Care Nurses. 2018;27(1):59-66.	Not reported	328 critical care nurses	Moral distress	0.71-0.84 for the composite α≥ 0.70 for all subscales
Gea-Caballero V, Castro-Sanchez E, Juarez- Vela R, Diaz-Herrera MA, de Miguel- Montoya I, Martinez-Riera JR. Essential elements of professional nursing environments in Primary Care and their influence on the quality of care. Enferm Clin. 2018;28(1):27-35.	Not reported	144 nurses	Evaluates the characteristics of nursing environments in primary care settings	No
Cho H, Han K. Associations Among Nursing Work Environment and Health-Promoting Behaviors of Nurses and Nursing Performance Quality: A Multilevel Modeling Approach. Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing. 2018.	57 units in 5 hospitals	432 nurses	Health promoting behaviors of hospital nurses	0.72-0.81 for the subscales
Al-Maaitah R, AbuAlRub RF, Al Blooshi S. Practice environment as perceived by nurses in acute care hospitals in Sharjah and North Emirates. Nursing forum. 2018;53(2):213- 222.	10 hospitals	450 nurses	Nurses' perceptions of their practice environment	0.90 for the composite
Akter N, Akkadechanunt T, Chontawan R, Klunklin A. Factors predicting quality of work life among nurses in tertiary-level hospitals, Bangladesh. Int Nurs Rev. 2018;65(2):182- 189	6 tertiary- level hospital	288 registered nurses	Level of quality of work life	0.90 for the composite
Zhang L, Wang A, Xie X, et al. Workplace violence against nurses: A cross-sectional study. Int J Nurs Stud. 2017;72:8-14	28 hospitals	3835 clinical nurses	Workplace violence	0.921 for the composite
Swiger PA, Raju D, Breckenridge-Sproat S, Patrician PA. Adaptation of the Practice Environment Scale for military nurses: a psychometric analysis. J Adv Nurs. 2017;73(9):2219-2236	42 US military treatment facilities	2608 nurses	Psychometric analysis	0.96 for the composite 0.81-0.90 for the subscales
Swiger PA, Patrician PA, Miltner RSS, Raju D, Breckenridge-Sproat S, Loan LA. The Practice Environment Scale of the Nursing Work Index: An updated review and recommendations for use. Int J Nurs Stud. 2017;74:76-84			46 articles published were reviewed in study	
Numminen O, Leino-Kilpi H, Isoaho H, Meretoja R. Development of Nurses' Professional Competence Early in Their Career: A Longitudinal Study. Journal of continuing education in nursing. 2017;48(1):29-39	Not reported	318 nurses	Examine competence development in nurses	0.77 to 0.86 for subscales (reports Lake, 2002)

Reference	# hospitals	# nurses	Outcome measure	Reliability (Cronbach's alpha)
Nantsupawat A, Kunaviktikul W, Nantsupawat R, Wichaikhum OA, Thienthong H, Poghosyan L. Effects of nurse work environment on job dissatisfaction, burnout, intention to leave. Int Nurs Rev. 2017;64(1):91-98	43 inpatient units in 5 university hospitals	1351 nurses	Association between work environment and nurse reported job dissatisfaction, burnout and intention to leave	0.85-0.91 for subscales (reports Nantsupawt et al, 2011)
Liu J, Zhou H, Yang X. Evaluation and Improvement of the Nurse Satisfactory Status in a Tertiary Hospital using the Professional Practice Environment Scale. Medical science monitor: international medical journal of experimental and clinical research. 2017;23:874-880	Not reported	1050 nurses	Associated factors influencing satisfaction	No
Hussein R, Everett B, Ramjan LM, Hu W, Salamonson Y. New graduate nurses' experiences in a clinical specialty: a follow up study of newcomer perceptions of transitional support. BMC Nurs. 2017;16:42	1 teaching hospital	87 new graduate nurses	Examine change in graduate nurses' perception	0.91 for the composite
Hallowell SG, Rogowski JA, Lake ET. How Nurse Work Environments Relate to the Presence of Parents in Neonatal Intensive Care. Advances in neonatal care : official journal of the National Association of Neonatal Nurses. 2017	104 US NICUs	6060 registered nurses	Infants whose parents were present during the NICU shift	No
Gasparino RC, Guirardello EB. Validation of the Practice Environment Scale to the Brazilian culture. J Nurs Manag. 2017;25(5):375-383	Not reported	209 nurses	Psychometric analysis of Brazilian version	0.86 for the composite 0.76-0.87 for the subscales
Elmi S, Hassankhani H, Abdollahzadeh F, Jafar Abadi MA, Scott J, Nahamin M. Validity and Reliability of the Persian Practice Environment Scale of Nursing Work Index. Iranian journal of nursing and midwifery research. 2017;22(2):106-111	Not reported	350 nurses	Psychometric analysis of Persian version	0.935 for the composite 0.70-0.92 for the subscales
Casalicchio G, Lesaffre E, Kuchenhoff H, Bruyneel L. Nonlinear Analysis to Detect if Excellent Nursing Work Environments Have Highest Well-Being. Journal of nursing scholarship: an official publication of Sigma Theta Tau International Honor Society of Nursing. 2017;49(5):537-547	2184 nursing units in 489 hospitals	33731 registered nurses	Burnout	No
Bruyneel L, Li B, Squires A, et al. Bayesian Multilevel MIMIC Modeling for Studying Measurement Invariance in Cross-group Comparisons. Med Care. 2017;55(4):e25-e35	87 nursing units in a single institution	87 nurse managers	Comparing and evaluating measurement invariance	No
Al-Hamdan Z, Manojlovich M, Tanima B. Jordanian Nursing Work Environments, Intent to Stay, and Job Satisfaction. Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing. 2017;49(1):103- 110.	Not reported	582 registered nurses	Intent to stay and job satisfaction	0.92 for the composite
Yokoyama M, Suzuki M, Takai Y, Igarashi A, Noguchi-Watanabe M, Yamamoto-Mitani N. Workplace bullying among nurses and their related factors in Japan: a cross-sectional survey. J Clin Nurs. 2016;25(17-18):2478- 2488	Not reported	825 nurses	Workplace bullying	0.75-0.84 for the subscsales
Schwendimann R, Dhaini S, Ausserhofer D, Engberg S, Zuniga F. Factors associated with high job satisfaction among care workers in Swiss nursing homes - a cross sectional survey study. BMC Nurs. 2016;15:37	162 nursing homes	4,145 care worker s	Job satisfaction	0.74-0.89 for subscales

Reference	# hospitals	# nurses	Outcome measure	Reliability (Cronbach's alpha)
Roche MA, Duffield C, Friedman S, Twigg D, Dimitrelis S, Rowbotham S. Changes to nurses' practice environment over time. J Nurs Manag. 2016;24(5):666-675	6 acute care hospitals	1605 nurses	To examine changes in the practice environment	0.82 for the composite 0.70-0.85 for the subscales
Hussein R, Everett B, Hu W, et al. Predictors of new graduate nurses' satisfaction with their transitional support programme. J Nurs Manag. 2016;24(3):319-326	Not reported	109 new graduate nurses	Satisfaction with transitional support program	0.91 for the composite
Gomez-Garcia T, Ruzafa-Martinez M, Fuentelsaz-Gallego C, et al. Nurses' sleep quality, work environment and quality of care in the Spanish National Health System: observational study among different shifts. BMJ Open. 2016;6(8):e012073	7 hospitals	635 registered nurses	Nurses sleep quality and quality of care	No
Duffield C, Roche M, Twigg D, Williams A, Clarke S. A protocol to assess the impact of adding nursing support workers to ward staffing. J Adv Nurs. 2016;72(9):2218-2225	20 pairs of matched wards	No	Protocol to asses the impact of adding nurse support workers	No
Brzyski P, Kozka M, Squires A, Brzostek T. How Factor Analysis Results May Change Due to Country Context. Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing. 2016;48(6):598-607	30 hospitals	2605 registered nurses	PES-NWI changes in the country context	0.72-0.89 for the subscales
Brooks-Carthon, J. M., Lasater, K. B., Rearden, J., Holland, S., & Sloane, D. M. (2016). Unmet nursing care linked to rehospitalizations among older Black AMI patients: A cross-sectional study of US hospitals. Medical care, 54(5), 457.	253 acute care hospitals	14879 registered nurses	Variable all-cause readmissions	No

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.

Other

If other: Survey tools are provided to nurses to complete themselves; most are done through electronic survey software, but the survey can be collected via paper.

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**.

Patient/family reported information (may be electronic or paper)

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).

This is nurse reported information. This measure is an eMeasure in the National Database of Nursing Quality Indicators, a voluntary benchmarking hospital network.

Maintenance of Endorsement (October 2018):

This measure is also an eMeasure in the Veterans Administration and the Military Hospital system

Here we support feasibility by presenting the numbers of hospitals, nursing units, and nurses, and response rates across years.

	NDNQI	VA					
	response rate	# hospitals	# units	#nurses	response rate	# hospitals	#nurses
2012	0.43	139	23,831				
2013	0.72	574	11,264	206,978	0.43	138	24,166
2014	0.68	395	7,557	131,619	0.47	141	28,930
2015	0.69	453	9,168	157,531	0.52	141	33,446
2016	0.72	349	8,236	132,764	0.53	141	35,700
2017	0.70	384	8,520	147,568	0.54	141	37,305
2018	0.56	141	38,967				

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

Attachment:

3c. Data Collection Strategy

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

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

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

In the Joint Commission testing project pilot test sites were given the option to collect the data via paper and pencil and enter data in the NSC tool, use the Survey Monkey tool created for the project, or share their data collected for NDNQI. Of the sites visited the majority used the survey monkey tool, followed by the NDNQI tool. One site loaded the tool into their Net-Learning intra-net program. Other large Nursing-Sensitive Care databases have used web-based tools and provide the link as well as a login for each nurse to allow for only one survey to be completed by each nurse.

Maintenance of Endorsement (October 2018):

The record of use of the measure by the NDNQI, the VA, and the military hospital systems demonstrates that there are minimal difficulties regarding data collection, availability of data, missing data (which was documented in the Measure Testing Submission Form submitted on 7/31/18, timing and frequency of data collection, sampling, nurse confidentiality, time and cost of data collection, or any other feasibility/implementation issues.

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 Maintenance of Endorsement (October 2018): The state of Colorado collects PES-NWI data from all hospitals with at least 100 beds every two years (odd years). These data are publicly reported www.cohospitalquality.org Professional Certification or Recognition Program American Nurses Credentialing Center Magnet Recognition Program https://www.nursingworld.org/organizational-programs/magnet/find-a-magnet-facility/ Quality Improvement (external benchmarking to organizations) National Database of Nursing Quality Indicators http://www.pressganey.com/solutions/clinical-quality/nursing-quality Quality Improvement (Internal to the specific organization) National Database of Nursing Quality Indicators http://www.pressganey.com/solutions/clinical-quality/nursing-quality

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

Centers for Medicare & Medicaid Services (CMS)
Hospital Inpatient Quality Reporting (Hospital IQR) program
Structural Measure: Participation in a Systematic Clinical Database Registry for Nursing Sensitive Care

URL:
http://qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228732621592
The trend in hospitals reporting to Medicare is that they participate in a Nursing-Sensitive Registry to comply with Medicare requirements:
For FY2011 (CY 2009):
1402 PPS providers participated in a Nursing sensitive registry
8 Non-PPS providers participated a Nursing sensitive registry
For FY2012 (CY 2010):
1491 PPS providers participated a Nursing sensitive registry
11 Non-PPS providers participated a Nursing sensitive registry
****
Many states have mandated collection and reporting of nursing-sensitive measures, for example:
Colorado: The Colorado Hospital Report Card
http://www.cohospitalquality.org/corda/dashboards/COLORADO_REPORT_CARD_BY_MEASURE/main.dashxml#cordaDash=1030
PES-NWI data reported:
2009 = 28 hospitals reported, range for Overall Composite Score 2.71 to 3.08
2011 = 29 hospitals reported, range for Overall Composite Score 2.69 to 3.25
Maintenance of Endorsement (October 2018):
2013 = 28 hospitals reported, range for Overall Composite Score 2.80 to 3.15
2015 = 26 hospitals reported, range for Overall Composite Score 2.78 to 3.10
2017 = 26 hospitals reported, range for Overall Composite Score 2.78 to 3.85
****
NDNQI (National Database of Nursing Quality Indicators, ANA): began in 1994, per NDNQI data is collected by more than 1500 hospitals nationwide. The annual RN survey is conducted in about half of the NDNQI hospitals. The PES-NWI was added to the annual RN survey in October 2006. Since its introduction, the number of hospitals that use of the PES-NWI in the National Database hospitals has increased on average 50% each year.
https://www.nursingquality.org/
NDNQI Annual RN Survey Data:
PES
Year Hospital Unit RNs
2006 97 1915 27255
2007 242 4845 81377
2008 330 6685 109100
2009 421 8532 142071
2010 524 10712 186566
2011 553 11513 206085
Maintenance of Endorsement (October 2018):
See 2012 – 2017 trend data presented above in section 3.b.2.
****
VANOD (Veterans Administration Nursing Outcomes Database): began development in 2002, this database includes data from all 153 VA facilities. The annual staff satisfaction survey includes the PES for RNs.
www.inqri.org/uploads/INQRIVANODPanel41309FINAL.ppt
Maintenance of Endorsement (October 2018):
See 2012 – 2018 trend data presented above in section 3.b.2.
****
The PES-NWI had been used in the military for the first time in 2002-2003 in a study of 22 Army Hospitals (see Patrician, Shang & Lake, 2010): Of the 1,793 surveys that were mailed, 955 were completed and returned,

representing an overall response rate of 53%, with a response range by hospital of 42–100%. Cronbach's alpha for the entire instrument was .94, with subscale alphas ranging from .82 to .87.
Patrician, P., Shang, J., & Lake, E. T. (2010). Organizational determinants of work outcomes and quality care ratings among army medical department registered nurses. Research in Nursing & Health, 33(2), 99-110.
The PES-NWI was also part of the nursing sensitive indicators used in the Military Nursing Outcomes Database (MilNOD) project from 2003-2006. Response rate from 13 military hospitals (Army, Navy and Air Force) during this time period was 35% overall. Cronbach´s alphas ranges from .77 to .85 for the subscales and .92-.93 for the composite score.
The PES-NWI was then used in 2010-2016 as part of the metrics for the Patient Caring Touch System (PCTS), a new nursing practice framework for the Army Nurse Corps. Every staff nurse in all Army (not Navy or Air Force) hospitals and clinics were surveyed in these years (except 2012 due to logistics issues). Cronbach´s alphas were .81 to .90 for the subscales and .96 for the composite. Response rates were variable by year.
The PES-NWI is being administered again in 2018 to all Army-affiliated staff nurses.

****
The PES-NWI is used internationally for quality improvement initiatives and research. There is great interest in using the survey in a variety of settings, the period 2004 to Spring 2012 includes 72 hospital administrators, 78 researchers, and 121 doctoral students, who notified the measure developer of use. Each year about 30 individuals seek advice and resources to use the PES. Over the eight year period, these inquiries have come from 34 states in the U.S. and 30 countries.
Maintenance of Endorsement (October 2018):
Since the 2012 endorsement, there have been, on average, 48 requests per year from researchers, hospital administrators, and PhD or master's students to use the instrument, including from 41 states, the District of Columbia, and 51 countries. These states and countries are displayed in the maps below. The countries represent all continents except Antarctica.
The PES-NWI has been translated into French (Swiss and Belgian variants), Spanish (Spain; Mexico due summer 2009), German (regular and Swiss variants), Japanese, Chinese, Korean, Dutch (Netherlands and Belgium) Russian, Armenian, Turkish, Portuguese (Brazilian only), Greek, Italian (Swiss variant), Finnish, Swedish, Polish, Flemish, and Arabic (due Summer 2009 via Jordan). In addition, validation of the UK English version is expected in summer 2009.

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?*)
n/a

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.

In the NDNQI, the VA, and the military hospitals, the performance results are shared in reports and dashboards with hospital managers to identify and address weaknesses in the nursing practice environments in their facilities. The number of facilities equaled 384 NDNQI hospitals in 2017 as well as all 141 VA and all 13 army hospitals nationally. The NDNQI is a national voluntary benchmarking database to track nursing quality

indicators. All VA and army hospitals collect the measure data. Interpretation is provided by NDNQI site coordinators in each hospital and at the VA and army hospitals by their quality and safety staff.

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.

Measure results are reported annually to the facilities that complete the survey. The data provided are descriptive statistics as well as trends for the subscales and the composite score.

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.

The NDNQI publishes monographs and holds conferences to provide exemplars of excellence to the facilities that collect, report, and evaluate the nursing-unit level data to assess the quality of nursing care and to transform the nursing units and improve outcomes. Press Ganey acquired NDNQI in 2015 and prepares annual strategic insight reports that are publicly available.

References to the monographs/reports:

Montalvo, I., & Dunton, N. (Eds.). (2007). Transforming nursing data into quality care: Profiles of quality improvement in US healthcare facilities. American Nurses Association.

Dunton, N., & Montalvo, I. (2009). Sustained improvement in nursing quality: Hospital performance on NDNQI indicators, 2007-2008. American Nurses Association.

Press Ganey Inc. 2017. Achieving Excellence: The Convergence of Safety, Quality, Experience and Caregiver Engagement http://healthcare.pressganey.com/2017-Strategic-Insights?s=White_Paper-PGPost

Example of a Conference:

Dunton, N., Staggs, V. & Potter, C. January 25, 2012. NDNQI Research Findings for the Advanced Site Coordinator. Preconference Workshop 002

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

The increasing trend in completion of the measure in the NDNQI membership indicates that the measure is valued by the member facilities. The inclusion of the measure in hospital dashboards indicates the measure is valued for monitoring quality.

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

The display of measure results in annual reports and manager dashboards demonstrates that the measure results are valuable to users.

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.

Typical feedback about the measure is that a reduction in length and testing for use in non-hospital settings is desired.

The measure has not been modified to date although reduction in survey length is planned for a future endorsement period.

Improvement

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

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

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

Maintenance of Endorsement (October 2018):

As reported in 1b, the use of the instrument by the NDNQI, the VA, and the military hospitals is for benchmarking and performance improvement.

The evidence presented above by Swiger et al (2017), shows that hospitals that have achieved Magnet recognition for meeting standards of nursing excellence have improved performance by having better work environments as compared to non-Magnet hospitals.

Additionally, this publication reports use of the instrument to improve nursing leadership, one subscale of the instrument:

Anderson, B. J., Manno, M., O´Connor, P., & Gallagher, E. (2010). Listening to nursing leaders: Using national database of nursing quality indicators data to study excellence in nursing leadership. Journal of Nursing Administration, 40(4), 182-187.

The article aims to examine nurse leadership qualities that create healthy work environments conducive to delivery of quality bedside care. The PES-NWI was used to assess qualities of exemplary nurse managers chosen by their staff. Researchers concluded that effective nurse leaders emphasized visibility, communication, and valued respect and empathy. These leadership strategies help to create healthy work environments that support nurse job satisfaction, nurse retention, and quality patient care delivery.

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.

As noted by Warshawsky and Havens, 2011 it is important that scoring and reporting of the PES-NWI be done consistently. There was inconsistency in reporting of subscales and composites across the many studies. There has also been variation in the unit of analysis for reporting, specifically nurse, nursing unit and organizational levels.

Maintenance of Endorsement (October 2018):

There have been no unexpected findings except for some non-significant results in some studies in the two new literature reviews. Nonsignificant results may be related to small sample sizes in some studies.

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

Maintenance of Endorsement (October 2018)

The unexpected benefits have been the worldwide use of the measure, generating comparable evidence to improve nursing work environments globally, and thereby improve patient safety and quality outcomes.

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.

No

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

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?

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

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.)

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:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): University of Pennsylvania, Center for Health Outcomes and Policy Research

Co.2 Point of Contact: Eileen, Lake, elake@nursing.upenn.edu, 215-898-2557-

Co.3 Measure Developer if different from Measure Steward:

Co.4 Point of Contact:

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.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2004

Ad.3 Month and Year of most recent revision:

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

Ad.5 When is the next scheduled review/update for this measure? 04, 2018

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: Regarding Ad.3: the measure has never been revised.