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 #: 0219
Measure Title: Post breast conservation surgery irradiation
Measure Steward: Commission on Cancer, American College of Surgeons
Brief Description of Measure: Percentage of female patients, age 18-69, who have their first diagnosis of breast cancer (epithelial malignancy), at AJCC stage I, II, or III, receiving breast conserving surgery who receive radiation therapy within 1 year (365 days) of diagnosis.
Developer Rationale: Improving the utilization of radiation with breast conservation surgery for breast cancer and optimizing risk of local recurrence.

Numerator Statement: Radiation therapy to the breast is initiated within 1 year (365 days) of the date of diagnosis
Denominator Statement: Include, if all of the following characteristics are identified:
Women
Age 18-69 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Primary tumors of the breast
Epithelial malignancy only,
AJCC Stage I, II, or III
Surgical treatment by breast conservation surgery (surgical excision less than mastectomy)
All or part of 1st course of treatment performed at the reporting facility
Known to be alive within 1 year (365 days) of diagnosis
Denominator Exclusions: Exclude, if any of the following characteristics are identified:
Men
Under age 18 at time of diagnosis
Over age 69 at time of diagnosis
Second or subsequent cancer diagnosis
Tumor not originating in the breast
Non-epithelial malignancies
Phyllodes tumor histology
Stage 0, in-situ tumor
Stage IV, metastatic tumor
None of 1st course therapy performed at reporting facility
Died within 12 months (365 days) of diagnosis
Patient participating in clinical trial that directly impacts delivery of the standard of care

Measure Type: Process
Data Source: Electronic Clinical Data : Registry, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 01, 2007 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 Form
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 *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.

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:

- National Comprehensive Cancer Network (NCCN) Practice Guidelines:
  - Radiation to the whole breast with or without boost (by photons, brachytherapy, or electron beam) to tumor bed. [Additional guidelines for node radiation based on extent of lymph node involvement]. **Level of evidence: Category 1**
  - Note t: Breast irradiation may be omitted in patients >=70 y of age with estrogen-receptor positive, clinically node negative, T1 tumors who receive adjuvant endocrine therapy. **Level of evidence: Category 1**
    - Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.
- Additional evidence included a systematic review of the body of evidence including multiple randomized clinical trials demonstrating approximately 75% reduction in risk of local recurrence with radiation and breast conservation - details on the total number of studies were not provided in previous submission form.
- The 2011 Committee expressed no concerns regarding the evidence underlying this measure.

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 updated the links for the guidelines and included the NCCN Categories of Evidence and Consensus – no changes were made to the evidence.

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as high-level evidence(Box 6) → Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on 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.

- The developer provided the following national trend data from the National Cancer Data Base (NCDB):

	2008	2012	2008-2012
# cases (numerator)	60,103	63,046	--
Mean performance Rate	88.1% (95% CI: 87.8-88.3)	90.7% (CI: 90.5-90.9)	--
IQR	84.0-97.0%	85.0-98.0%	--
Range	--	--	0-98.0%

- The developer stated that more recent performance data was not provided because all adjuvant therapy information are likely incomplete for the most recent year until programs have had time to collect this information.
- In 2011, the Committee noted that there was a moderate gap in performance for this measure.

Disparities:

Race/ethnicity

2012	Non-Hispanic white	Non-Hispanic black	Hispanic	Asian/Hawaiian/Pacific Island
# of patients	47,161	6,305	2,886	1,567
mean performance rates (MPR)	92.1%	86.4%	83.1%	89.7%

Age

2012	Age 18-49	Age 50-59	Age 60-69
# of patients	14,054	22,415	--
mean performance rates (MPR)	89.8%	91.4%	--

Insurance Status

2012	Private Insurance	Medicare	Medicaid/No insurance	Other government insurance
# of patients	--	11,576	5,391	--
mean performance rates (MPR)	91.7%	89.7%	86.1.%	89.6%

- The developer provided additional disparities data on median household income, residents without a high school degree, and facility type.
- In 2011, the Committee noted that there were demonstrated disparities on the basis of age, race/ethnicity, and other factors.

Questions for the Committee:

- *Without more recent performance data, does the Committee agree there is a gap in care that warrants a national performance measure?*

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Evidence provided applies directly to the process outcome such that higher use of the intervention results in 75 % reduction in local recurrence.

This is a process measure. It directly relates to the need to deliver radiation in the adjuvant setting following breast conserving therapy. No new evidence was presented since this was a maintenance submission. It is unlikely that any new data is available. The evidence is of high quality, high quantity, and consistent. There is broad consensus with the data. As previously presented, a systematic review of the body of evidence including multiple randomized clinical trials demonstrating approximately 75% reduction in risk of local recurrence with radiation and breast conservation. I rate this MODERATE according to the algorithm.

1b. Performance Gap

Comments:

Moderate performance gap demonstrated in less than optimal performance (although improving) of 90.7%. Recent performance data not presented but 2011 data indicated variability in care and disparities in race/ethnicity, age, insurance status, education, and SES.

No new data presented. The developer stated that more recent performance data was not provided because all adjuvant therapy information is likely incomplete for the most recent year until programs have had time to collect this information. Previous data by subgroups was presented. Disparities in care were noted in most areas. Previously, a moderate gap in performance was demonstrated. The target for this measure should be high since the complete treatment plan includes radiation post-operatively. Therefore it remains important to measure. Without new data to demonstrate an ongoing performance gap, I rate this moderate.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): registry and paper medical records. This is not an eMeasure.

Specifications:

- This is a facility-level measure.
- The numerator is defined as radiation therapy to the breast is initiated within 1 year (365 days) of the date of diagnosis.
- The denominator includes nine data elements:
  - women
  - age 18-69 at time of diagnosis
  - known or assumed to be first or only cancer diagnosis
  - primary tumors of the breast
  - epithelial malignancy only
  - AJCC Stage I, II, or III
  - surgical treatment by breast conservation surgery (surgical excision less than mastectomy)

- all or part of 1st course of treatment performed at the reporting facility
- known to be alive within 1 year (365 days) of diagnosis

- Denominator exclusions include:
  - men
  - under age 18 at time of diagnosis
  - over age 69 at time of diagnosis
  - second or subsequent cancer diagnosis
  - tumor not originating in the breast
  - non-epithelial malignancies
  - Phyllodes tumor histology
  - Stage 0, in-situ tumor
  - Stage IV, metastatic tumor
  - none of 1st course therapy performed at reporting facility
  - died within 12 months (365 days) of diagnosis
  - patient participating in clinical trial that directly impacts delivery of the standard of care – this is an update since last endorsement date
- A calculation algorithm is provided.
- All cases which meet the measure criteria are included in the denominator. If a required data element is missing the case is flagged for additional review.
- Diagnosis codes are based on the Facility Oncology Registry Data Standards (FORDS), which were revised in 2016. Therefore, no ICD-9 or ICD-10 codes are provided for this measure.
- The database is a hospital cancer registry reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 1,400 cancer programs and approximately 55,700 cases from all CoC-accredited cancer programs. The mean performance rates across all CoC-accredited cancer programs in 2007 was 84.1 and 84.7 in 2008. Cancer programs in the 75th percentile had performance rates of 97.2 and 98.3 in each respective year; 8.6% of programs had statistically low outlier performance rates (<52%), SD=22.3%.

Describe any updates to testing:

- The developer provided updated performance rates from 2012 for all CoC-accredited cancer programs:
  - Mean performance rate: 90.7%
  - 10th percentile: 81.8% and less
  - Range: 0-98% (2008-2012)

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the

individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Overall performance rates do not meet criterion.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empirical testing as specified (Box 2) → empirical validity testing at patient level (Box 3)→use rating from validity testing of patient-level data elements (Box 10)→percent agreement provided for one data element (Box 11)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Questions for the Committee:
- *See questions under Validity*

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Randomly selected charts were reviewed by site surveyors to determine completeness and validity of data reported to registry. The measure denominator and numerator were viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Describe any updates to testing: The developer provided additional details on data element validity testing - see below

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
- The developer provided the following information about the dataset: Survey sites and data collection occurred in 2009 and 2010. In 2009, 391 sites were reviewed and 5,712 charts - 5,390 of these charts were breast cases from 2006; representing 15.8% of measure eligible cases. In 2010, 423 sites were reviewed and 6,752 charts –

6,370 of these charts were breast cases from 2007; representing 15.7% of measure eligible cases.

- Data elements reviewed:
  - confirmation of timing of adjuvant therapy
  - documentation of treatment recommended but not received
  - assessment of missing and incomplete tumor characteristics

Validity testing results:

- The developer provided the following testing results:
  - "Assessment of timing for radiation therapy for cases in which treatment was administered significantly early for this measure (<=60 days after diagnosis) had high concordance with 91.4% in 2006 and 92.2% in 2007. A total of 494 cases in which treatment was not coded as having been administered during both 2006 and 2007 diagnoses, there was 59% agreement with missing radiation therapy."
- The developer provided percentage agreement results for two of the data elements included in the numerator (timing of radiation therapy and therapy recommended but not received). NQF guidance states that testing should be done for all critical data elements.
- Site surveyors determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.
- Developers provided only percentage agreement statistics. While these were fairly high for radiation therapy, a percentage agreement of 59% for missing radiation therapy is concerning; no additional results were provided (e.g., kappa scores, which indicate agreement over and above chance; sensitivity or specificity statistics).

Questions for the Committee:

- *Does the measure adequately identify and include breast cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients are excluded from the measure for the following reasons:
  - Men
  - Under age 18 at time of diagnosis
  - Over age 69 at time of diagnosis
  - Second or subsequent cancer diagnosis
  - Tumor not originating in the breast
  - Non-epithelial malignancies
  - Phyllodes tumor histology
  - Stage 0, in-situ tumor
  - Stage IV, metastatic tumor
  - None of 1st course therapy performed at reporting facility
  - Died within 12 months (365 days) of diagnosis
  - Patient participating in clinical trial that directly impacts delivery of the standard of care
- Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure.
- In 2012-2013, 1 case (<0.01%) was excluded due to patient participating in clinical trial that directly impacts delivery of the standard of care; this exclusion does not affect estimated performance rates for this measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions of sufficient frequency and variation across providers to be needed (and outweigh the data*

collection burden)?

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Performance data is presented above under opportunity for improvement. Complete details of the data is presented in 1b.

Question for the Committee:

- *Given the data provided in 1b., does the measure identify meaningful differences about quality across facilities?*

2b6. Comparability of data sources/methods:

- Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b7. Missing Data

- The developer describes in S.22 that all cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review. The developer does not provide information on the frequency of missing data or potential impact on results.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) →validity testing conducted with patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Data elements are clearly defined, measure widely implemented, minimal concerns that measure can be consistently implemented.

The data elements are clearly defined with appropriate codes and descriptors. I have no concerns about the ability to implement this measure.

No concerns.

The specifications are consistent with the evidence and are valid.

2a2. Reliability Testing

Comments:

Validity testing results not provided for all critical elements and percentage agreement results are low.

The reliability algorithm suggests that the testing is insufficiently reliable and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff.

2b2. Validity Testing

Comments:

Only percentage agreement statistics provided with only a percentage agreement of 59%.

The validity algorithm suggests that the testing is insufficiently valid and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff.
At this point, the measure does not pass the reliability and validity standards as set forth by NQF.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Exclusions are consistent with evidence. 2b4: None. The measure does identify meaningful differences. Validity testing only conducted for two numerator data elements.
I identify no threats to validity. As with all abstracted data, the potential for abstraction errors are significant. The missing data that was presented with this submission is of concern. They report approximately 59% missing data and describe their follow up plan but provide no further information regarding success for obtaining data. Exclusions are appropriate. There is no risk adjustment. Meaningful differences should be detectable.

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.

- Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry); some data elements are in defined fields in electronic sources.
- Data collection burden due to manual chart abstraction from paper medical records.

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

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Dependent upon structure of medical record with lower feasibility in paper records, all data may not easily be automatically extracted in all EHR's and reliance on CTRs presents some level of burden for smaller organizations and non COC accredited institutions.
All of the data is routinely used during patient care and should be available in the medical record. The majority of the data requires abstraction which is fraught with potential area and burdensome to obtain. I have no concerns about the data collection strategy. I rate this MODERATE.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- **Public Reporting:** The Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.
- **Regulatory and Accreditation Programs**: The Commission on Cancer (CoC) is a consortium of professional organizations dedicated to improving survival and quality of life for cancer patients through standard-setting, prevention, research, education, and the monitoring of comprehensive quality care. One of the standards for CoC-accredited cancer programs requires program meet established estimated performance rates with accountability measures or develop an action plan for improvement. Approximately 1500 cancer programs are CoC-accredited constituting nearly 70% of diagnosed cases.
- **Quality Improvement with Benchmarking**: The National Cancer Data Base (NCDB) provides a benefit for CoC-accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP). CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer. This application is available to over 1500 CoC-accredited cancer programs. CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

Improvement results:

- The overall facility level compliance rates have increased from 88.1% (95% CI: 87.8-88.3) in 2008 to 90.1% (90.5-90.9) IQR=85-98% n=63,046 in 2012 representing a 2.3% overall improvement in quality.
- Across all patient demographics progress towards the goal of quality improvement for this measure was noted between 2008-2012; i.e. race/ethnicity, age groups, insurance status, income levels, educational levels, cancer program types, and census regions.

Unexpected findings (positive or negative) during implementation

- This measure, as specified, is susceptible to under-reporting of the adjuvant RT component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. However, CoC accredited programs have demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. Additionally, the CoC Standards require direct review and oversight of this measure compliance be monitored by an attending physician (Cancer Liaison Physician, CLP) on staff at the center on a quarterly basis.

Potential harms

- Developer did not identify any unintended consequences related to this measure.

Feedback :

- Developer did not identify any specific feedback loops related to this measure.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Public reporting through the Pennsylvania Health Care Quality Alliance and the CoC National data base.

Publically reported and in use in several settings. Since this measure is associated with evidence to suggest a substantial improvement in outcome if radiation is omitted, it is important to improve health care delivery. I rate this HIGH.

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0219	**NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

Process

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Directly applicable - randomized trials examining the measure specified treatment

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** Multiple randomized clinical trials

1c.6 **Quality of Body of Evidence** *(Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* High quality evidence

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Strong level of consistency

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

Approximate 75% reduction in risk of local recurrence with radiation and breast conservation

1c.9 **Grading of Strength/Quality of the Body of Evidence.** Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN); Early Breast Cancer Trialists Collaborative Group

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

NCCN Categories of Evidence and Consensus

Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

All recommendations are category 2A unless otherwise noted.

http://www.nccn.org/professionals/physician_gls/categories_of_consensus.asp

1c.13 **Grade Assigned to the Body of Evidence:** Level 1

1c.14 **Summary of Controversy/Contradictory Evidence:** None

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

See 1b.4

1c.16 **Quote verbatim, <u>the specific guideline recommendation</u>** *(Including guideline # and/or page #)***:**

Page BINV-2

Radiation to the whole breast with or without boost (by photons, brachytherapy, or electron beam) to tumor bed (category 1). [Additional guidelines for node radiation based on extent of lymph node involvement].

Note t: Breast irradiation may be omitted in patients >=70 y of age with estrogen-receptor positive, clinically node negative, T1 tumors who receive adjuvant endocrine therapy (category 1)

1c.17 **Clinical Practice Guideline Citation:** NCCN Clinical Practice Guidelines v1.2016

1c.18 **National Guideline Clearinghouse or other URL:**

http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

1c.19 **Grading of Strength of Guideline Recommendation.** Has the recommendation been graded? Yes

1c.20 **If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN)

1c.21 **System Used for Grading the Strength of Guideline Recommendation:** Other

1c.22 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

1c.23 **Grade Assigned to the Recommendation:** Level I

1c.24 **Rationale for Using this Guideline Over Others:** All guidelines recommend radiation after breast conservation surgery

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency**:** High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0219_Evidence_MSF5.0_Data.doc,BCSRT_0225_Evidence_2016-635932238997751583.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Improving the utilization of radiation with breast conservation surgery for breast cancer and optimizing risk of local recurrence.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The nationally recognized National Cancer Data Base (NCDB), jointly sponsored by the American College of Surgeons and the American Cancer Society, is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70 percent of newly diagnosed cancer cases nationwide and 30 million historical records. Data from the NCDB were analyzed.

The NCDB collects data from CoC accredited cancer programs on an annual basis; the data we collect is in accordance with standard registry procedures. In January of 2015, 2013 diagnoses were collected. This information was released to accredited cancer programs in the late summer. However, we find information on some of the therapies which take longer to be received are not complete when submitted, therefore the Commission on Cancer does not begin surveying or holding programs accountable for their Estimated Performance Rates (EPRs) until the year after it is released to ensure adequate adjuvant therapy information has been documented. We generally see a decrease in compliance for the most recent year until programs have had time to collect this information, since we don't feel the EPRs are accurate at initial release we did not include the 2013 data in the application for this measure and used the next most recent annual rate of 2012 for this measure.

The mean performance rate (MPR) for this measure has increased from 88.1% (95% CI: 87.8-88.3) IQR=84-97% n= 60,103 in 2008 to 90.7% (90.5-90.9) IQR=85-98% n=63,046 in 2012 representing a 2.3% overall improvement in quality. The minimum hospital performance rate is 0% with a 98% maximum in all years assessed 2008-2012.

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.**

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

Race/ethnicity

Race/ethnicity was defined as non-Hispanic white, non-Hispanic black, Hispanic, Asian/Hawaiian/Pacific Island or other race/ethnicity. Non-hispanic whites had the highest mean performance rates (MPR) in 2012 at 92.1% (95% CI: 91.9-92.4) n=47,161 in 2013, followed by Asian Pacific Islanders 89.7% (88.3-91.1) n=1,567, Blacks 86.4% (85.6 -87.1) n=6,305, and Hispanics 83.1% (81.9-84.4) n=2,886. Between 2008 and 2012, MPR increased in all ethnic groups. In 2012, the white population's MPR was significantly higher than any of the other race/ethnicities ($p<0.05$). There is no significant difference in the MPR of the Pacific Islander and the Black population. The Hispanic population's MPR is significantly below all of the other race/ethnic groups during both 2008 and 2012 ($p< 0.05$).

Age

Age groups were defined as, 18-49, 50-59, 60-69. Since 2008, each age group saw a gains in performance with the measure. In 2012, patients, age 50-59, had the highest MPR (91.4%; 95% CI: 91.0-91.8; n-22,415), a rate that was significantly higher when compared to the under 50 (89.8%; 95% CI: 89.3-90.3; n=14,054).

Insurance Status

Insurance status is defined as insurance at the time of diagnosis and stratified into private, Medicare, Medicaid/No insurance. From 2008 through 2012, patients with each insurance type saw a gain in performance. In 2012, uninsured and Medicaid patients had the lowest MPR at 86.1.% (95% CI: 85.2-87.0) n=5391, Medicare at 89.7% (89.2-87.490.2) n=11,576, Other Government 89.6% (87.5-91.7), with private insurance having the highest performance rates at 91.7% (91.4-91.9).

SES – Median household income within zip code

Income quintiles at the zip code level were assessed based on the 2012 American Community Survey. Patients that resided in communities with a median income of <$36,000 annually at diagnosis experienced lower performance in 2012 than patients from

communities with a median income above $36,000. That mean performance rate (MPR) was significantly lower (87.1%, 95% CI: 86.3-87.9; n=6,676)(p< 0.05). Those that resided in a community with and income of $36,000-$43,000, were significantly different from the remaining categories [MPR of 90.0% (95% CI: 89.4-90.6; n=9,503)]. The remaining communities whose median income was above $43,999 were not significantly different and ranged in a MPR of 91.3%; (95% CI: 90.8-91.8 to 91.8%; n=11,567) to 91.5% (95% CI: 91.1-91.9; n=19,542).

SES- Proportion of residents without a high school degree within zip code
Patients that resided in communities at time of diagnosis with the lowest proportion of no high school degree (<7%) had higher rates of performance in 2012 92.6% (92.2-93.0) n=18,591) than patients from communities with the highest proportion of patients with no high school degree (>21%) 86.3 (85.6-87.0; n=8,854) Likewise, the performance increase from 2008 was smaller for patients from communities with the highest proportion of no high school degree (>21%) 2.9% gain.

Facility Type
Facility type was assessed by CoC-accreditation status; facility types include Comprehensive Community Cancer Programs, Integrated Network Cancer Programs, Community Cancer Programs and by Teaching/Research programs. In 2012, only patients treated at the smaller community cancer programs experienced a significant difference in the MPR (88.6%; 95% CI: 87.9-89.4, n=7,223). However, the MPR for these smaller hospital increased some 2.8% between 2008 and 2012. For the remaining cancer program types, there were not significant differences in the MPR: Academic/research (AR), 90.6%, CI: 90.2—91.0, n=19,951); Comprehensive community cancer programs (CCCP), 91.3%, CI: 91.0-91.6, n=30,132); and Integrated network cancer programs (ICNP), 90.8,%, CI: 90.0-91.6, n=4,931). All of these latter hospitals experienced changes in performance from 2008-20012 ranging from 1.8% CCCPs, 2.1%, for INCPs, and 5.6% for ARs.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
There is extensive evidence from randomized clinical trials demonstrating the impact of radiation with breast conservation surgery. It reduces the risk of local recurrence in the breast and may have a small impact on survival. The limitation for the purpose of a measure for provider accountability to women under the age of 70 is because of high level evidence that women with small, estrogen receptor positive cancer (the majority of women over age 70 with breast cancer) gain only a very small reduction in local recurrence and no difference in life-time mastectomy rate and no difference in survival. The impact as measured by performance gap, improvement in outcome and numbers of cases affected has been specifically examined by Hasset et al.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Early Breast Cancer Trialists Collaborative Group (EBCTCG) et al. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised tirals. Lancet 2011;378(9804):1707-1716. 2. Hassett MJ, Hughes ME, Niland JC, et al. Selecting high priority quality measures for breast cancer quality improvement. Med Care 2008;46:762-770. 3. Hughes KS, Schnaper LA, Berry D, et al. Lumpectomy plus tamoxifen with or without irraditaion in women 70 years of age or older with early breast cancer. New Engl J Med 2004;351:971-977.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Breast

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Disparities

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.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

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)
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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Since the last endorsement minor changes have been made to this measure. The word considered has been replaced in the numerator statement with recommended to be more consistent with the registry codes used to assess this measure. These include: the removal of rare histologies (Malignant phyllodes tumors, 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor, 8980 - Carcinosarcoma, 8981 - Carcinosarcoma, embryonal) ,and phyllodes tumors from inclusion in the denominator based on lack of evidence to support inclusion. Users are also allowed to exclude cases from the denominator based on patient enrollment in a clinical trial that directly impacts delivery of the standard of care.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Radiation therapy to the breast is initiated within 1 year (365 days) of the date of diagnosis

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
1 year (365 days)

S.6. **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, 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.
Regional Treatment Modality [NAACCR Item#1570]=20-98, and Date Radiation Started [NAACCR Item#1210] <= 365 days following the Date of Diagnosis [NAACCR Item#340]

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Include, if all of the following characteristics are identified:
Women
Age 18-69 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Primary tumors of the breast
Epithelial malignancy only,
AJCC Stage I, II, or III
Surgical treatment by breast conservation surgery (surgical excision less than mastectomy)
All or part of 1st course of treatment performed at the reporting facility
Known to be alive within 1 year (365 days) of diagnosis

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Sex [NAACCR Item#220]=2; Age at Diagnosis [NAACCR Item#230] < 70; AND Surgical Procedure of the Primary Site [NAACCR Item#1290] = 20–24

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude, if any of the following characteristics are identified:
Men
Under age 18 at time of diagnosis
Over age 69 at time of diagnosis
Second or subsequent cancer diagnosis
Tumor not originating in the breast
Non-epithelial malignancies
Phyllodes tumor histology
Stage 0, in-situ tumor
Stage IV, metastatic tumor
None of 1st course therapy performed at reporting facility
Died within 12 months (365 days) of diagnosis
Patient participating in clinical trial that directly impacts delivery of the standard of care

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No stratification applied

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
This measure score is calculated by dividing the numerator cases by denominator eligible cases.
Denominator eligible cases are assessed in a step-wise fashion:
- Include breast cancer cases
- Exclude Patients enrolled in a clinical trial that directly impacts delivery of the standard of care
- Include female cases only
- Include first and only primary tumors
- Include epithelial tumors staged according to AJCC 7th edition
- Exclude 8940 - Mixed tumor, malignant, NOS; 8950 - Mullerian mixed tumor; 8980 – Carcinosarcoma; 8981 - Carcinosarcoma, embryonal
- Include invasive tumors only
- Exclude pathologic evidence of in situ or metastatic disease
- Exclude clinical evidence of in situ or metastatic disease
- Include cases with all or part of the first course of treatment was performed at the reporting facility
- Include cases with receipt of breast conserving surgery
- Include patient reported living within 365 days from the date of diagnosis

Numerator cases are then assessed from denominator eligible cases:
- Cases are included in the numerator if
Radiation therapy was administered within 365 days following diagnosis.

The measure score is calculated with the numerator divided by the denominator.

See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
No sampling

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on*

minimum response rate.)
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
All cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospital cancer registry data, reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0219_MeasureTesting_MSF5.0_Data.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0219 **NQF Project:** Cancer Project

2. RELIABILITY & 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. (**evaluation criteria**)

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

This measure has been implemented by the ACoS CoC since 2007 across all CoC-accredited cancer programs, and reports on approximately 55,700 cases per year to almost 1,400 cancer programs.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

Cancer registry case records reported to the NCDB are reviewed annually, annualized hospital performance rates are provided back to CoC accredited cancer programs via the CoC´s Cancer Program Practice Profile Report (CP3R) using the denominator and numerator criteria documented in response to items 2a1.3 and 2a1.7, respectively, in the Specifications section. (http://www.facs.org/cancer/ncdb/cp3r.html)

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

The mean performance rates across all CoC-accredited cancer programs was 84.1 in 2007 and 84.7 in 2008. The two years available at the time of this writing. Cancer programs in the 75th percentile had performance rates of 97.2 and 98.3 in each respective year. Even with high aggregate performance rates demonstrated by programs room for improvement across the system of CoC-accredited programs remains, with 8.6% of programs with statistically low outlier performance rates (<52%), SD=22.3%.

The mean performance rates in all CoC-accredited cancer programs was 90.7% in 2012. Cancer programs in the 10th percentile had a performance rate of 81.8% and less. The minimum hospital performance rate is 0% with a 98% maximum in all years assessed 2008-2012.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

See 2a2.1. This measure has been implemented across all CoC-accredited cancer programs and subject to local review by standing committees of these hospitals and site surveyors at the time of accreditation site visits.

During Commission on Cancer Survey Site visits in 2009 and 2010, surveyors validated not more than 25 charts.

During 2009 – 391 accredited sites were reviewed, including 5,712 charts. This included an average of 14 charts per survey (IQR 6-22). 5,390 of these charts were breast cases; representing 15.8% of measure eligible cases.
During 2010- 423 accredited sites were reviewed, including 6,752 charts. This was based on an average of 14 charts per survey (IQR 6 – 22). 6370 of these charts were breast cases; representing 15.7% of measure eligible cases

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

Performance rates are reviewed and discussed, randomly selected charts are reviewed by the site surveyor to ascertain the completeness and validity of the data recorded in the local cancer registry and reported to the NCDB and included in the CP3R reporting application.

Major areas of review completed by site surveyors included but were not limited to, confirmation of timing of adjuvant therapy, documentation of treatment recommended but not received, assessment of missing and incomplete tumor characteristics.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure has a high degree of user acceptability, the measure denominator and numerator are viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Assessment of timing for radiation therapy for cases in which treatment was administered significantly early for this measure (<=60 days after diagnosis) had the high concordance with 91.4% in 2006 and 92.2% for 2007 cases. A total of 494 cases in which treatment was not coded as having been administered during both 2006 and 2007 diagnoses, there was 59% agreement with missing radiation therapy.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

The NCDB collects all diagnosed cases within cancer programs. The measure exclusions as described are the opposite of the measure inclusion criteria. Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure.

The exception to this is the exclusion of, Patient participation in a clinical trial which directly impacts the standard of care.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed. For all cases applicable to this measure, in 2012 -2013, 1 case was excluded from the measure denominator based on the exclusion of patient participation in a clinical trial which directly impacts the standard of care.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

Measure exclusions were used for n=1 (<0.01%) and does not affect estimated performance rates for this measure.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of*

data; if a sample, characteristics of the entities included):

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

Differences in data performance was described in performance gaps.

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:

2c. Disparities in Care: H☐ M☐ L☐ I☐ NA☐ (*If applicable, the measure specifications allow identification of*

disparities.)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*:

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

This measure was not specified to report stratified performance rates, however the CoC´s recently released (2011) "real clinical time" Rapid Quality Reporting System (RQRS) (http://www.facs.org/cancer/ncdb/rqrs.html) reports back measure-specific performance rates by a number of strata, eg. patient age, sex, ethnicity, insurance status, and area-based SES. RQRS hosts a prospective treatment alert system, and so performance rates are both high and consistent with clinical expectation, however room for potential improvement remains. In a comparative analysis of 16 NCI/NCCCP pilot sites using RQRS with a comparative group of 25 other CoC-accredited cancer programs also using RQRS revealed that at NCCCP cancer programs white patients more frequently received post-op RT (91.4%) than did African-American women (87.4%) following breast conservation surgery; and Medicaid recipients less frequently (84.8%) received post-op RT than insured patients (91.6%). Comparative rates from the 25 non-NCCCP programs were slightly lower across the board, however the patterns of disparate receipt of care were mirrored in this group of hospitals. Analysis from cases diagnosed 2008-2010.

Additional disparities data was presented in section 1.b. of this application.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

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)

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)*

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.**
The ACoS/CoC implementation of this measure is framed around the feasibility of data collection and reporting considerations. Cancer registries in the United States depend on a multitude of information sources in order to completely abstract case records and be in compliance with State, Federal and private sector accreditation requirements. Commission on Cancer Standards require case abstracting to be performed by a Certified Tumor Registrars (CTRs). CTRs must pass an exam and maintain continuing education. In the past decade, great strides have been made within the cancer registration community in terms of electronic capture of registry data from electronic pathology systems and electronic health records. However, until EHR systems are universally implemented in the US and fully integrated within hospital-level cancer registry systems, registry data will depend upon some level of human review and intervention to ensure data are complete and accurately recorded. Robust data quality edits are applied to the data at all levels of cancer data abstraction and processing. These edits standardize coded information and ensure its accuracy.

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.**
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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
1)The infrastructure to monitor compliance with this measure has been in place since 2005 to assess and feed-back to approximately 1,500 Commission on Cancer (CoC) accredited centers performance rates for this measure. CoC accredited cancer programs account for 70-80% of patients affected by this measure. This measure is currently reported to CoC accredited programs through the National Cancer Data Base (NCDB) using the Cancer Program Practice Profile Report (CP3R) web-based audit and feed-back reporting tool. The CP3R is generally described at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. In addition, this measure is also reported to over 100 cancer programs participating in its "real clinical time" feedback reporting tool through its Rapid Quality Response System (RQRS). An overview of the RQRS is available at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Both of these reporting tools have been utilized in the cancer registry community and will not produce an undue burden on the data collection network.

2)The data for this measure are key elements already collected in all hospital registries. This measure has been reviewed using cancer registry data. The CoC data demonstrates variation in the measure. Registries have demonstrated the ability to identify gaps in data collection and to correctly identify therapy in the majority of cases. The measure is readily implemented.

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)*.

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Pennsylvania Health Care Quality Alliance http://www.phcqa.org/ Regulatory and Accreditation Programs Commission on Cancer https://www.facs.org/quality-programs/cancer/coc Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Commission on Cancer, National Cancer Data Base https://www.facs.org/quality-programs/cancer/ncdb

4a.1. For each CURRENT use, checked above, provide:

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

a) Public Reporting
Pennsylvania Health Care Quality Alliance
Purpose: The Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania.
Area: The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.

d) Regulatory and Accreditation Programs
Commission on Cancer
Purpose: The Commission on Cancer (CoC) is a consortium of professional organizations dedicated to improving survival and quality of life for cancer patients through standard-setting, prevention, research, education, and the monitoring of comprehensive quality care.
One of the standards for CoC-accredited cancer programs requires program meet established estimated performance rates with accountability measures or develop an action plan for improvement. Standards Manual: https://www.facs.org/quality-programs/cancer/coc/standards
Approximately 1500 cancer programs are CoC-accredited constituting nearly 70% of diagnosed cases.

f) Quality Improvement with Benchmarking
Commission on Cancer, National Cancer Data Base
Purpose: The National Cancer Data Base (NCDB) provides a benefit for CoC-accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP).

CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer and is described https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. This application is available to over 1500 CoC-accredited cancer programs

CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cqip

RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers - See more at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

2008 88.1 (87.8 – 88.3); n=60,103
2009 89.7 (89.4 – 89.9), n=59,784
2010 91.5 (91.3 – 91.7); n=60,930
2011 91.9 (91.7 – 92.1); n=64,026
2012 90.7 (90.5 – 90.9); n=63,043

The overall facility level compliance rates have increased from 881% (95% CI: 87.8-88.3) in 2008 to 90.1% (90.5-90.9) IQR=85-98% n=63,046 in 2012 representing a 2.3% overall improvement in quality.

Across all patient demographics progress towards the goal of quality improvement for this measure was noted between 2008-2012; i.e. race/ethnicity, age groups, insurance status, income levels, educational levels, cancer program types, and census regions. The highest mean performance rates were experienced by non-Hispanic Whites, patients 50-59 years old, patients with private insurance, communities with higher incomes and more education, and patients living the Midwest Census region. The data demonstrate a need to continue to use this measure to address issues of disparity that analyses of the data demonstrate. Arguably, one could make the case that the change is better data capture. However, the significant differences found across patient demographics demonstrates that disparity exists. In particular, a focus on the Hispanic population, the younger age groups, the uninsured and Medicaid patients, communities with lower incomes and education levels, and patients living in the South and West Census Regions. These differences in mean performance rates are described in 1b.2 and 1b.4

Census region
There were differences in MPR across census regions in 2012. In order from highest MPR (Midwest) to the lowest (South), the following MPRs, 95% CIs, and n are reported: Midwest (93.9%, CI: 93.5-94.3, n=15,715); West (92.0%, CI: 91.1-93.0, n=3,047); Northeast (90.5%, CI: 90.1-91.0, n=14,780); Pacific (89.6%, CI: 88.9-90.3, n=7,862); South (88.8%, CI: 88.4-89.2, n= 21,491)

4b.2. **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.**

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
This measure, as specified, is susceptible to under-reporting of the adjuvant RT component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. However, CoC accredited programs have demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. Additionally, the CoC Standards require direct review and oversight of this measure compliance be monitored by an attending physician (Cancer Liaison Physician, CLP) on staff at the center on a quarterly basis.

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
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 completely harmonized?

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.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Commission on Cancer, American College of Surgeons
Co.2 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-
Co.3 **Measure Developer if different from Measure Steward:** Commission on Cancer, American College of Surgeons
Co.4 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-

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 Developers: Christopher Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); David Winchester, MD, FACS (Northshore University Health System, Evanston IL); Diana Dickson-Witmer, MD, FACS (Chistiana Health Care System, Wilmington DE); Kelly Hunt, MD, FACS (MD Anderson Cancer Center, Houston TX); Marilyn Leitch, MD, FACS (University of Texas – Southwestern, Dallas TX); Katherine Virgo, PhD (American Cancer Society)

The current Measure workgroup includes:
Charles Cheng MD, FACS (Fox Valley Surgical Associates, Appleton, WI), Daniel McKellar, MD, FACS (Wayne Healthcare, Greenville, OH), David Jason Bentrem, MD (Northwestern Memorial Hospital, Chicago, IL),
Karl Bilimoria, MD, FACS (Northwestern Univ/Feinberg Sch of Med, Chicago, IL), Lawrence Shulman MD (University of Pennsylvania, Philadelphia, PA), Matthew A Facktor, MD FACS (Geisinger Medical Center, Danville, PA), Ted James (University of Vermont, Burlington, VT)

This panel meets at least once annually to review quality measures currently supported and implemented by the ACoS Commission on Caner and to investigate and consider/review development of possible new measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 05, 2007
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?** 11, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 0220
Measure Title: Adjuvant hormonal therapy
Measure Steward: Commission on Cancer, American College of Surgeons
Brief Description of Measure: Percentage of female patients, age >18 at diagnosis, who have their first diagnosis of breast cancer (epithelial malignancy), at AJCC stage T1cN0M0,IB to III, who´s primary tumor is progesterone or estrogen receptor positive with tamoxifen or third generation aromatase inhibitor (recommended or administered) within 1 year (365 days) of diagnosis.
Developer Rationale: Improve the utilization of hormone therapy for women with estrogen receptor positive breast cancer.

Numerator Statement: Hormone therapy is administered within 1 year (365 days) of the date of diagnosis or it is recommended but not received
Denominator Statement: Include if all of the following characteristics are identified:
Women
Age >=18 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Epithelial malignancy only
Primary tumors of the breast
AJCC T1cN0M0 or Stage IB - III
Primary tumor is estrogen receptor positive or progesterone receptor positive
All or part of 1st course of treatment performed at the reporting facility
Known to be alive within 1 year (365 days) of date of diagnosis
Denominator Exclusions: Exclude, if any of the following characteristics are identified:
Men
Under age 18 at time of diagnosis
Second or subsequent cancer diagnosis
Tumor not originating in the breast
Non-epithelial malignancies, exclude malignant phyllodes tumors, 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor , 8980 - Carcinosarcoma,8981 - Carcinosarcoma, embryona
Stage 0, in-situ tumor
AJCC T1mic, or T1a tumor
Stage IV, metastatic tumor
Primary tumor is estrogen receptor negative and progesterone receptor negative
None of 1st course therapy performed at reporting facility
Died within 1 year (365 days) of diagnosis,
Patient enrolled in a clinical trial that directly impacts delivery of the standard of care

Measure Type: Process
Data Source: Electronic Clinical Data : Registry, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 01, 2007 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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:

- National Comprehensive Cancer Network (NCCN) Practice Guidelines:
  - Systemic Adjuvant treatment – hormone receptor positive- HER2- Positive: (Page BINV-5): pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant endocrine therapy +/- adjuvant chemotherapy with trastzumab. **Level of evidence: Category 1 (**Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate)
  - Systemic Adjuvant treatment – hormone receptor positive- HER2- Negative: (PageBINV-6): pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant endocrine therapy +/- adjuvant chemotherapy
- Additional evidence included a systematic review of the body of evidence including multiple randomized clinical trials and meta-analysis demonstrating 25% reduction in risk of distant cancer recurrence and death - details on the total number of studies were not provided in previous submission form.
- The 2011 Committee expressed no concerns regarding the evidence underlying this measure

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 updated the links for the guidelines and included the NCCN Categories of Evidence and Consensus – no changes were made to the evidence.

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as high-level evidence (Box 6) → Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on 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.

- The developer provided the following national trend data from the National Cancer Data Base (NCDB):

	2008	2012
# of hospitals	1,487	1,436
# patients	74,017	85,570
Hospital Estimated Performance Rates (EPR)	78.7%	85.5%
Standard Deviation	23.5	19.8
IQR	72.3-94.4%	84-96.9%
Range	0-100%	0-100%

  - The developer stated that more recent performance data was not provided because all adjuvant therapy information are likely incomplete for the most recent year until programs have had time to collect this information.
  - In 2011, the Committee noted that there was a moderate gap in performance for this measure.

Disparities:

Race/ethnicity

2012	Non-Hispanic white	Non-Hispanic black	Hispanic	Asian/Hawaiian/Pacific Island	Other
# of patients	65,386	8,468	4,927	2,399	4,390
Estimated performance rates (EPR)	89.3%	82.5%	80.1%	85.7%	84.6

Age

2012	Age 18-49	Age 50-59	Age 60-69
# of patients	20,444	18,943	18,515
Estimated performance rates (EPR)	85.9%	79.8%	81.9%

Insurance Status

2012	Private Insurance	Medicare	Medicaid/No insurance	Other government insurance
# of patients	48,273	28,550	6,839	901

Estimated performance rates (EPR)	88.4%	88.3%	81.9%	86.0%

- The developer provided additional disparities data on income and education.
- In 2012, the Committee noted that disparities were demonstrated between African American and white females.

Questions for the Committee:

- *Without more recent performance data, does the Committee agree there is a gap in care that warrants a national performance measure?*

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The high level evidence (systematic review and multiple randomized clinical trials and meta-analyses) for this process measure is directly applicable to the process and there is strong consistency in the evidence.

Process measure that directly applies to the delivery of adjuvant therapy. A systematic review of the body of evidence including multiple randomized clinical trials and meta-analysis demonstrating 25% reduction in risk of distant cancer recurrence and death demonstrating the significance of this measure. There are many high quality, consistent studies to demonstrate the benefit of adjuvant chemotherapy with long-term follow-up. No new evidence was presented but there is no need for new evidence to support this measure. According to the algorithm, this evidence is rated MODERATE.

**This is a process measure that addresses the percentage of female patients, aged 18 or older at diagnosis of an epithelial breast cancer, tumor ER/PR+, and Hormone therapy (Tamoxifen or AI) recommended or administered within 1 year (365 days) of date of diagnosis. The evidence provided consists of SRs that include RCTs. Clinical Practice Guideline (e.g., NCCN). No changes in evidence have been provided by the developer. **

1b. Performance Gap

Comments:

Performance data on the rate of use of adjuvant hormonal therapy was provided and demonstrates a gap in care with variability present in the data and less than optimum use. Multiple disparities race/ethnicity, insurance status, and income.

The developer stated that more recent performance data was not provided because all adjuvant therapy information is likely incomplete for the most recent year until programs have had time to collect this information. During the last submission, Hospital Estimated Performance Rates (EPR) were 78.7% in 2008 and 85.5% in 2012 suggesting improvement but an ongoing gap. Subgroup analysis demonstrated a disparity performance gap in most subgroups. Affects large populations. Without new evidence of ongoing performance gap, I rate this MODERATE.

**This measure does indicate a gap in care and disparities are noted.

Numerator = hormone therapy administered with 365 days or recommended, but not received.

Denominator = eligibility criteria (inclusion - most exclusion is simply the opposite of inclusion so not true exclusion criteria - except clinical trial that impacts SOC. Improvement has been demonstrated from 2008 to 2012, disparity noted primarily with non-hispanic black, hispanic populations as compared to non-hispanic white (EPR for white=89.3, Non-hispanic black = 82.5, and hispanic = 801. Improvement is also needed with patients receiving Medicaid versue private insurance or Medicare.**

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): registry and paper medical records. This is not an eMeasure.

Specifications:

- This is a facility-level measure.
- The numerator is defined as hormone therapy that is administered within 1 year (365 days) of the date of diagnosis or it is recommended but not received. Reasons the hormone therapy was recommended but not administered include:
  - Contraindicated due to patient risk factors
  - Patient died prior to planned or recommended therapy
  - Recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record.
  - Recommended by the patient´s physician, but treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record.
- The denominator includes:
  - Women
  - Age >=18 at time of diagnosis
  - Known or assumed to be first or only cancer diagnosis
  - Epithelial malignancy only
  - Primary tumors of the breast
  - AJCC T1cN0M0 or Stage IB - III
  - Primary tumor is estrogen receptor positive or progesterone receptor positive
  - All or part of 1st course of treatment performed at the reporting facility
  - Known to be alive within 1 year (365 days) of date of diagnosis
- Denominator exclusions include:
  - Men
  - Under age 18 at time of diagnosis
  - Second or subsequent cancer diagnosis
  - Tumor not originating in the breast
  - Non-epithelial malignancies, exclude malignant phyllodes tumors, 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor , 8980 - Carcinosarcoma,8981 - Carcinosarcoma, embryona
  - Stage 0, in-situ tumor
  - AJCC T1mic, or T1a tumor
  - Stage IV, metastatic tumor
  - Primary tumor is estrogen receptor negative and progesterone receptor negative
  - None of 1st course therapy performed at reporting facility
  - Died within 1 year (365 days) of diagnosis,
  - Patient enrolled in a clinical trial that directly impacts delivery of the standard of care– this is an update since last endorsement date
- A calculation algorithm is not provided in measure submission form.
- All cases which meet the measure criteria are included in the denominator. If a required data element is missing the case is flagged for additional review.
- Diagnosis codes are based on the Facility Oncology Registry Data Standards (FORDS), which were revised in 2016. Therefore, no ICD-9 or ICD-10 codes are provided for this measure.
- The database is a hospital cancer registry reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base.
- In 2012, the Committee recommended that in future iterations, the measure capture that the patients are receiving the appropriate dose of hormonal therapy, appropriateness of hormonal therapy based upon menopausal state of the patient, and patient adherence to the hormonal therapy through filled prescriptions.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 1,400 cancer programs and approximately 65,200 cases from from all CoC-accredited cancer programs. The mean performance rates across all CoC-accredited cancer programs in 2007 was 76.6 and 77.1 in 2008. Cancer programs in the 75th percentile had performance rates of 95.8 and 96.9 in each respective year; 3.5% of programs had statistically low outlier performance rates (<15%). The SD of the distribution of performance rates for this measure is noticably greater than that of the other measures, in excess of 27%.

Describe any updates to testing:

- The developer provided updated performance rates from 2012 for all CoC-accredited cancer programs:
  - #of facilities: 1,436 facilities
  - Mean hospital level estimated performance rate (EPR): 85.5
  - SD: 19.8
  - Range: 0-100
  - IQR: 84.0-96.9
- The developer stated that there were 156 programs with EPRs in the lowest 10th percentile of 81.0%.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Overall performance rates do not meet criterion.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empirical testing as specified (Box 2) → empirical validity testing at patient level (Box 3)→use rating from validity testing of patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Questions for the Committee:

- *See questions under Validity*

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Randomly selected charts were reviewed by site surveyors to determine completeness and validity of data reported to registry. The measure denominator and numerator were viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Describe any updates to testing: The developer provided additional details on data element validity testing - see below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer provided the following information about the dataset: Survey sites and data collection occurred in 2009 and 2010. In 2009, 391 sites were reviewed and 5,712 charts - 5,390 of these charts were breast cases from 2006; representing 15.8% of measure eligible cases. In 2010, 423 sites were reviewed and 6,752 charts – 6,370 of these charts were breast cases from 2007; representing 15.7% of measure eligible cases.
- Data elements reviewed:
  - confirmation of timing of adjuvant therapy
  - documentation of treatment recommended but not received
  - assessment of missing and incomplete tumor characteristics

Validity testing results:

- The developer provided the following testing results:
  - "Assessment of timing for hormone therapy for cases in which treatment was administered significantly early (<=60 days after diagnosis) for this measure had the highest concordance with 84.3 in 2006 and 79.1% for 2007 cases. There was 77.9% and 91.1% agreement in 2006 and 2007 diagnoses respectively for hormone therapy which was recommended but not administered for this measure. A total of 298 cases with missing hormone receptor status were reviewed, this information was found in nearly 90% of these cases."
- The developer provided percentage agreement results for two of the data elements included in the numerator (timing of hormone therapy and therapy recommended but not received). NQF guidance states that testing should be done for all critical data elements.
- Site surveyors determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.
- Developers provided only percentage agreement statistics which indicated a decrease of 5.2% for the data

element timing of hormone therapy from 2006 (84.3%) to 2007 (79.1%); no additional results were provided (e.g., kappa scores, which indicate agreement over and above chance; sensitivity or specificity statistics).

Questions for the Committee:

- *Does the measure adequately identify and include colon cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

Exclude, if any of the following characteristics are identified:

- Men
- Under age 18 at time of diagnosis
- Second or subsequent cancer diagnosis
- Tumor not originating in the breast
- Non-epithelial malignancies, exclude malignant phyllodes tumors, 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor , 8980 - Carcinosarcoma,8981 - Carcinosarcoma, embryona
- Stage 0, in-situ tumor
- AJCC T1mic, or T1a tumor
- Stage IV, metastatic tumor
- Primary tumor is estrogen receptor negative and progesterone receptor negative
- None of 1st course therapy performed at reporting facility
- Died within 1 year (365 days) of diagnosis,
- Patient enrolled in a clinical trial that directly impacts delivery of the standard of care

- The measure exclusions as described are the opposite of the measure inclusion criteria. The cases excluded are those in which the clinical evidence does not support inclusion in the quality measure.
- In 2012-2013, 5 cases (<0.01%) were excluded due to patient participating in clinical trial that directly impacts delivery of the standard of care; this exclusion does not affect estimated performance rates for this measure.
- In 2012, the Committee questioned why there was no exclusion for pregnancy or planned pregnancy. The developer noted that of the 110,000 women reported on, 63 had a secondary diagnosis code with pregnancy. This equates to one half of one percent. Half of these women did ultimately receive hormonal therapy; it is plausible that those women received the therapy after delivery. Consequently, the number of patients excluded for pregnancy would be extremely minimal. With respect to planned pregnancy, it is not feasible to ascertain planned pregnancy with respect to the measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Performance data is presented above under opportunity for improvement. Complete details of the data are presented in 1b.

Question for the Committee:

- *Given the data provided in 1b., does the measure identify meaningful differences about quality across facilities?*

2b6. Comparability of data sources/methods:

- Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b7. Missing Data

- The developer describes in S.22 that all cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review. The developer does not provide information on the frequency of missing data or potential impact on results.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) → validity testing conducted with patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The measure data elements are well defined.

Data elements clearly defined. Codes are provided. I have no concerns that this measure can be consistently implemented.

This is a facility-level measure. The numerator and denominator are described. Numerator includes "contraindicated due to patient risk factors" Can tihs data element be easily extracted - are risk factors relevant to this measurement clear? It is not clear if the 2012 Committees recommendation to capture the hormonal dose was included. Also, menopausal status and patient adhrerence was recommended. Not clear if those were included this time and if not, ratonale provided.

Specifications consistent with the evidence presented.

The specifications are consistent with the evidence.

The measure specifications are consistent with the data.

2a2. Reliability Testing

Comments:

Does not meet reliability testing requirements, insufficient number of data elements in numerator assessed for percent agreement.

The reliability algorithm suggests that the testing is insufficiently reliable and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff.

Reliability testing requirements were insufficien. A calculation algorithm was not provided.

2b2. Validity Testing

Comments:

Scope of testing was adequate and completed appropriately.

**The validity algorithm suggests that the testing is insufficiently valid and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff. At this point, the measure does not pass the reliability and validity standards as set forth by NQF. **

**Since reliability was not sufficient, validity testing will count. However, validity is viewed to be insufficient - did not have agreement testing on all critical data elements (only two).

Randomly selected charts were reviewed - tested against gold standard. Kappa scores or sensitivey/specificity were not used.**

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Exclusions are consistent with the evidence and no groups are inappropriately excluded. Evidence for support of use indicates that this measure does identify meaningful difference about quality.
Exclusions are appropriate and consistent with the evidence. No risk adjustment. This measure when implemented should be able to detect meaningful differences and should provide comparable results. Missing data is a threat. The developers discuss a strategy to identify and address missing data but no follow up is provided.
No information as to the impact of missing data on results.

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.

- Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry); some data elements are in defined fields in electronic sources.
- Data collection burden due to manual chart abstraction from paper medical records.

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

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Data generally captured by EHR although variability exists and collection may rely on CTRs creating some level of burden.
All of the data is routinely used during patient care and should be available in the medical record. The majority of the data requires abstraction which is fraught with potential area and burdensome to obtain. I have no concerns about the data collection strategy. I rate this MODERATE.
Extracted by trained staff (e.g., registry personnel). Some of the data are in defined fields in elecrronic sources and others are via paper medical records (manual extraction).

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- The Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.
- PPS-Exempt Cancer Hospital Quality Reporting program: In 2010 the Affordable Care Act required the Centers for Medicare and Medicaid Services (CMS) to establish a specialized quality reporting program for the PPS-exempt cancer hospitals. The resulting PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program measures allow consumers to compare the quality of care given at the eleven PPS-exempt cancer hospitals currently participating in the program. This includes 11 PPS-Exempt Cancer Hospitals.
- The Commission on Cancer (CoC) is a consortium of professional organizations dedicated to improving survival and quality of life for cancer patients through standard-setting, prevention, research, education, and the monitoring of comprehensive quality care. One of the standards for CoC-accredited cancer programs requires program meet established estimated performance rates with accountability measures or develop an action plan for improvement. Approximately 1500 cancer programs are CoC-accredited constituting nearly 70% of diagnosed cases.
- The National Cancer Data Base (NCDB) provides a benefit for CoC-accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP). CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer. This application is available to over 1500 CoC-accredited cancer programs. CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers.
- Quality Oncology Practice Initiative: In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.
- QOPI® Certification Program: The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score. Please see a description of the QOPI® program above for details.

Improvement results:

- At the hospital level, the mean EPR in 2008 was 78.7%. The EPR in 2012 was 85.5.
- EPRs increased in all census regions between 2008 and 2012, with a 10% increase in the Northeast region, a 9% increase in the West and Pacific regions, a 7% increase in the South region, and a 6% increase in the Midwest region.

Unexpected findings (positive or negative) during implementation:

- This measure, as specified, is susceptible to under-reporting of the adjuvant hormone therapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. However, CoC accredited programs have demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. It does take additional time to collect and report this adjuvant therapy information. Additionally, the CoC´s Program Standards require review of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

Potential harms

- Developer did not identify any unintended consequences related to this measure.

Feedback :

- Developer did not identify any specific feedback loops related to this measure.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

The measure is being publicly reported by the Pennsylvania Health Care Quality Alliance, the Commission on Cancer. Performance results further the goal of high-quality efficient healthcare through reduction of the risk of distant cancer recurrence.

The measure is publically reported and used in multiple settings. This is an important measure that can be used to improve outcomes and the quality of health care. No unintended consequences. Benefits of implementing this measure are high.

Used by several entities for accountability and performance improvement: e.g., CoC, QOPI, PHCQA.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0387 : Oncology: Hormonal therapy for stage IC through IIIC, ER/PR positive breast cancer

Harmonization

- The measures are not harmonized because these assess different levels of analysis and different data systems are used to determine eligibility and compliance. 0387 assesses whether hormone therapy was prescribed whereas 0220 assesses whether hormone therapy was administered within one year of diagnosis or if it was recommended but not received based on patient refusal, medical co-morbidity or other valid reasons.
- 0220 assesses compliance at the facility level while 0387 assesses individual physician or practice level performance and the measures use different data sources as well.

Pre-meeting public and member comments

-

NQF #: 0220 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three sub criteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
Process

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
Directly applicable - randomized trials examining this process

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: Multiple randomized clinical trial and meta-analysis

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* High level evidence

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Strong consistency

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
Approximate 25% reduction in risk of distant cancer recurrence and death

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN): Early Breast Cancer Trialists Collaborative Group

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

1c.13 **Grade Assigned to the Body of Evidence:** Level I

1c.14 **Summary of Controversy/Contradictory Evidence:** None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
See 1b.4

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
"Adjuvant endocrine therapy" [depending on other tumor characteristics also includes "+/- adjuvant chemotherapy" or "+ adjuvant chemotherapy"]

Systemic Adjuvant treatment – hormone receptor positive- HER2- Positive: (Page BINV-5)
pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant endocrine therapy +/- adjuvant chemotherapy with trastzumab (category 1)

Systemic Adjuvant treatment – hormone receptor positive- HER2- Negative: (PageBINV-6)
pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant endocrine therapy +/- adjuvant chemotherapy

1c.17 Clinical Practice Guideline Citation: NCCN Clinical Practice Guidelines v1.2016

1c.18 National Guideline Clearinghouse or other URL: http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: National Comprehensive Cancer Network (NCCN)

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: Level I, IIA, IIB, III

1c.23 Grade Assigned to the Recommendation: Level I

1c.24 Rationale for Using this Guideline Over Others: All guidelines recommend hormone therapy with hormone receptor positive breast cancer

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0220_Evidence_MSF5.0_Data.doc,0HT_220_Evidence_2016-635953596415814634.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Improve the utilization of hormone therapy for women with estrogen receptor positive breast cancer.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The nationally recognized National Cancer Data Base (NCDB), jointly sponsored by the American College of Surgeons and the American Cancer Society, is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70 percent of newly diagnosed cancer cases nationwide and 30 million historical records. Data from the NCDB was analyzed.

The NCDB collects data from CoC accredited cancer programs on an annual basis; the data we collect is in accordance with standard registry procedures. In January of 2015, 2013 diagnoses were collected. This information was released to accredited cancer programs in the late summer. However, we find information on some of the therapies which take longer to be received are not complete upon initial submission and need time to document receipt of adjuvant therapy. Therefore the CoC does not begin surveying or holding programs accountable for their Estimated Performance Rates (EPRs) until the year after it is released to ensure adequate adjuvant therapy information has been documented. We generally see a slight decrease in compliance for the most recent year until programs have had time to collect this information, since we don't feel all adjuvant therapy information are complete at initial submission we did not include the 2013 data in the application for this measure and used the next most recent annual rate of 2012 for this measure.

There were 74,017 patients eligible for this measure in 2008 and 85,570 eligible in 2012. There were 1,487 facilities that had patients eligible for this measure in 2008, and 1,436 facilities with eligible patients in 2012. The EPR increased 7% between 2008 and 2012. For EPRs at the hospital level, the mean EPR in 2008 was 78.7%, Standard Deviation 23.5, minimum=0, maximum=100%, interquartile range 72.3-94.4. The mean hospital level EPR in 2012 was 85.5, Standard deviation=19.8, minimum=0, maximum=100, interquartile range: 84.0-96.9.

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.**

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
The data source is described in 1b.1. Demographic comparisons by race/ethnicity, age, insurance status, census tract income/education, and census region follow.

Race/Ethnicity: EPRs for Non-Hispanic Whites, Blacks, Asians, and Hispanics, and other races were generated. Between 2008 and

2012, EPRs increased in all race ethnic groups. Increases from 2008 to 2012 were 7% for Non-Hispanic Whites, 9% for Non-Hispanic Blacks, 13% for Hispanics, 8% for Asians, and 7% for Other. EPRs (%) in 2012 by race/ethnicity show the lowest EPRs for Hispanics (80.1; 95% CI: 79.0-81.2; n=4,927), and the highest EPRs among Non-Hispanic Whites (89.3; 95% CI: 89.1-89.5, n=65,386). Non-Hispanic blacks had the 2nd lowest EPRs (82.5; 95% CI: 81.7-83.3; n=8,468). Asian and other race/ethnicity groups had the second highest EPRs (Asian: 85.7; 95% CI: 84.3-87.1; n=2,399), (Other: 84.6; 95% CI: 83.5-85.6, n=4,390). EPRs in 2008 also show the lowest EPRs for Hispanics (67.3; 95% CI: 65.8-68.8, n=3,837), followed by Non-Hispanic Blacks (73.9: 95% CI: 72.8-74.9, n=6,422), Asians (77.9; 95% CI: 75.9-79.8, n=1,797), Other (77.9; 95% CI: 76.9-78.8, n=7,693), and Non-Hispanic Whites (82.0; 95% CI: 81.7-82.3, n=54,286).

Age: Age group comparisons include 18-49, 50-59, 60-69, 70-79, and 80 years and over. Between 2008 and 2012, EPRs increased in all age groups. Percent increases were 8% in ages 18-49, 50-59, and 79 and over, and 7% in ages 60-69 and 70-79. By age in 2012, the highest EPR was found in those 70-79 (89.1; 95% CI: 88.6-89.6, n=13,804), followed by ages 60-69 (88.9; 95% CI: 88.5-89.3, n=23,063), ages 50-59 (87.9; 95% CI: 87.5-88.4, n=22,410), ages > 79 (86.1; 95% CI: 85.2-87.0, n=5,849), and ages 18-49 (85.9; 95% CI: 85.4-86.4, n=20,444). In 2008, the highest EPR is found for ages 70-79 (82.5; 95% CI: 81.8-83.1, n=11,661), followed by ages 60-69 (81.9 95% CI: 81.4-82.5, n=18,515), ages 50-59 (79.8; 95% CI: 79.3-80.4, n=18,943), age 80 and over (78.1; 95% CI: 77.0-79.1, n=5,930); ages 18-49 (77.4; 95% CI: 76.8-78.0,n=18,968).

Insurance: Insurance category comparisons include Private insurance, Medicare, Medicaid/Not Insured, Other government, and Other/Unknown. EPRs increased in all insurance groups between 2008 and 2012. Percent increases were 8% for Medicaid/Not Insured and Private Insurance, 7% for Medicare, 9% for Other government, and 11% for Other/Unknown. EPRs in 2012 for insurance status show the highest rates of compliance for Private insurance (88.4; 95% CI: 88.1-88.7, n=48,273), followed by Medicare (88.3; 88.0-88.7, n=28,550), other government (86.0; 95% CI: 83.8-88.3, n=901), Medicaid/Not Insured (81.9; 95% CI: 81.0-82.8, n=6,839), and Other/Unknown (80.5, 95% CI: 78.1-83.0, n=1,007 EPRs in 2008 show highest rates of compliance for Medicare (81.5; 95% CI: 81.0-82.0, n=23,948), followed by Private insurance (80.1; 95% CI: 79.7-80.5, n=43,725), Other government (77.0; 95% CI: 73.7-80.2, n=634), Not insured/Medicaid (74.1; 95% CI: 72.8-75.3, n=4,750), and Other/Unknown (69.9; 95% CI: 67.0-72.8, n=960).

Income: By quintiles of median income in 2012 based on census tract, the highest EPRs are found in the top three tiers of median income. EPRs increased for all income groups between 2008 and 2012. Percent increases between 2008 and 2012 were 8% in the lowest quintile, 6% in the 20-40% quintile, 7% in the 40-60% and 60-80% quintiles, and 9% in the highest quintile. The EPRs are: lowest quintile (84.9; 95% CI: 84.2-85.6, n=9,486); 20-40% Quintile (86.6; 95% CI: 86.1-87.2, n=13,742); 40-60% quintile (88.5; 95% CI: 88.0-89.0, n=16,155); 60-80% quintile (88.9; 95% CI: 88.5-89.3, n=21,189); highest Quintile (88.1; 87.7-88.5,n=24,718). In 2008, the lowest EPR is found in the lowest quintile of median income. The EPRs are: lowest quintile (76.7; 95% CI: 75.8-77.6, n=8,186); 20-40% quintile (80.5; 95% CI: 79.8-81.2, n=11,835); 40-60% quintile (81.3; 80.6-81.9, n=13,941); 60-80% quintile (81.8; 95% CI: 81.2-82.4, n=17,792); and the highest quintile (79.1; 95% CI: 78.6-79.7, n=21,211).

Education: Percent of census tracts with no High School degree by quartiles are used to compare EPRs. EPRs increased in all education groups between 2008 and 2012. Percent increases were 8% in the 21% or higher, 13-20%, and less than 7% no high school degree groups, and 7% in the 7-12% no High School degree group. EPRs by percent with no high school degree from 2012 census data show the highest EPRs in census tracts with the lowest percentage of No High School Degree. EPRs by education are as follows: 21% or more no High School degree (83.0 95% CI: 82.4-83.7, n=12,633); 13-20% no High School Degree (87.2; 95% CI: 86.7-87.6, n=20,408); 7-13% no High School degree (88.6; 95% CI: 88.2-89.0, n=); less than 7% no High School Degree (89.8; 95% CI: 89.4-90.2, n=24,121). EPRs in 2008 are as follows: 21% or more no high school degree (74.8; 95% CI: 74.0-75.6, n=10,779; 13-20% no high school degree (79.7; 95% CI: 79.1-80.3, n=17,026), 7-12% no high school degree (81.4; 95% CI: 80.9-81.8, n=24,521), and < 7% no high school degree (81.9; 95% CI: 81.3-82.4, n=20,677).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
NA

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
There is extensive evidence that hormone (endocrine) therapy with hormone receptor positive breast cancer reduces the risk of local recurrence, contralateral breast cancer, distant recurrence, and death. Measures specifies use of tamoxifen or 3rd generation aromatase inhibitor rather than specifying tamoxifen for pre-menopausal and aromatase inhibitor for post-menopausal because a) Difficulty in clearly identify from records or administrative data the menopause status and b) variation in appropriate use of tamoxifen in post-menopausal, and some reasonable use of aromatase inhibitor in pre-menopausal women with the use of ovarian suppression.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Early Breast Cancer Trialists Collaborative Group (EBCTCG) et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 2011;378(9793):771:784. 2. Cuzick J, Sestak I, Baum M, et al. Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 10-year analysis of the ATAC trail. Lancet Oncol 2010;11:1135-1141. 3. Burstein JH, Prestrud AA, Seidenfeld J, et al. American Society of Clinical Oncology clinical practice guidelines: update on adjuvant endocrine therapy for women with hormone receptor positive breast cancer. J Clin Oncol 2010;28:3784-3796.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Breast

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Disparities

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.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

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)
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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Since the last endorsement, the minor changes have been made to the measure specifications:
These include: the removal of rare histologies
Malignant phyllodes tumors,
8940 - Mixed tumor, malignant, NOS
8950 - Mullerian mixed tumor
8980 - Carcinosarcoma
8981 - Carcinosarcoma, embryonal
from inclusion in the denominator based on lack of evidence to support inclusion.

Cases are now able to be excluded from the measure denominator if they are participating in a clinical trial which directly impacts the standard of care.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Hormone therapy is administered within 1 year (365 days) of the date of diagnosis or it is recommended but not received

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
1 year (365 days)

S.6. **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, 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.

Hormone Therapy recommended and not received [NAACCR Item#1400]=82-87 (82:not recommended/administered because it was contraindicated due to patient risk factors, 85:not administered because the patient died prior to planned or recommended therapy,86:It was recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record. 87: it was recommended by the patient´s physician, but this treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record)

OR; Hormone Therapy administered [NAACCR Item#1400]=1, AND Date Hormone Therapy Started (NAACCR Item#710] <=365 days following Date of Diagnosis [NAACCR Item# 340]

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*

Include if all of the following characteristics are identified:
Women
Age >=18 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Epithelial malignancy only
Primary tumors of the breast
AJCC T1cN0M0 or Stage IB - III
Primary tumor is estrogen receptor positive or progesterone receptor positive
All or part of 1st course of treatment performed at the reporting facility
Known to be alive within 1 year (365 days) of date of diagnosis

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*

Sex [NAACCR Item#220]=2; and
Age [NAACCR Item# 230] >=18; and
Stageable Epithelial tumors histology [NAACCR Item# 522] 8000-8576, 8941-8949 and
Invasive tumor behavior [NAACCR Item# 522] =3 and
AJCC T1c or Stage IB-III:Tumor Size [NAACCR Item#2800]= 11-989, 992-995 and AJCC pN [NAACCR Item#890]=0, I-, I+, 0M-, M=, 0M+ OR AJCC pN [NAACCR Item#890]=1,1M, 1M1, 1A, 1B, 1C,2, 2A, 2B, 3, 3A, 3B, or 3C; and
CS SSF1 (ERA) [NAACCR Item#2880]=010 or 030; AND CS SSF2 (PRA) [NAACCR Item#2890]=010 or 030;
AND Surgical Procedure of the Primary Site [NAACCR Item#1290] = 20–90

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Exclude, if any of the following characteristics are identified:
Men
Under age 18 at time of diagnosis
Second or subsequent cancer diagnosis
Tumor not originating in the breast
Non-epithelial malignancies, exclude malignant phyllodes tumors, 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor , 8980 - Carcinosarcoma,8981 - Carcinosarcoma, embryona
Stage 0, in-situ tumor
AJCC T1mic, or T1a tumor
Stage IV, metastatic tumor
Primary tumor is estrogen receptor negative and progesterone receptor negative
None of 1st course therapy performed at reporting facility
Died within 1 year (365 days) of diagnosis,
Patient enrolled in a clinical trial that directly impacts delivery of the standard of care

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*

See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No stratification applied

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
All cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.24.
Electronic Clinical Data : Registry, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospital cancer registry data, reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0220_MeasureTesting_MSF5.0_Data.doc,0220_MeasureTesting_HT_04012016-635953795775862982.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0220	**NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

This measure has been implemented by the ACoS CoC since 2007 across all CoC-accredited cancer programs, and reports on approximately 65,200 cases per year to almost 1,400 cancer programs.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

Cancer registry case records reported to the NCDB are reviewed annually, annualized hospital performance rates are provided back to CoC accredited cancer programs via the CoC´s Cancer Program Practice Profile Report (CP3R) using the denominator and numerator criteria documented in response to items 2a1.3 and 2a1.7, respectively, in the Specifications section. (http://www.facs.org/cancer/ncdb/cp3r.html)

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

The mean performance rates across all CoC-accredited cancer programs was 76.6 in 2007 and 77.1 in 2008. The two years available at the time of this writing. Cancer programs in the 75th percentile had performance rates of 95.8 and 96.9 in each respective year. Even with high aggregate performance rates demonstrated by programs room for improvement across the system of CoC-accredited programs remains, with 3.5% of programs with statistically low outlier performance rates (<15%). The SD of the distribution of performance rates for this measure is noticeably greater than that of the other measures, in excess of 27%.

There were 1,436 facilities with eligible patients for this measure in 2012. The mean hospital level EPR in 2012 was 85.5, Standard deviation=19.8, minimum=0, maximum=100, interquartile range: 84.0-96.9. In 2012 there were 156 programs with EPRs in the lowest 10th percentile of 81.0%.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

See 2a2.1. This measure has been implemented across all CoC-accredited cancer programs and subject to local review by standing committees of these hospitals and site surveyors at the time of accreditation site visits.

During Commission on Cancer Survey Site visits in 2009 and 2010, surveyors validated not more than 25 charts.

During 2009 – 391 accredited sites were reviewed, including 5,712 charts. This included an average of 14 charts per survey (IQR 6-22). 5,390 of these charts were breast cases; representing 15.8% of measure eligible cases.

During 2010- 423 accredited sites were reviewed, including 6,752 charts. This was based on an average of 14 charts per survey (IQR 6 – 22). 6370 of these charts were breast cases; representing 15.7% of measure eligible cases

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

Performance rates are reviewed and discussed, randomly selected charts are reviewed by the site surveyor to ascertain the completeness and validity of the data recorded in the local cancer registry and reported to the NCDB and included in the CP3R reporting application.

Major areas of review completed by site surveyors included but were not limited to, confirmation of timing of adjuvant therapy, documentation of treatment recommended but not received, assessment of missing and incomplete tumor characteristics.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure has a high degree of user acceptability, the measure denominator and numerator are viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Assessment of timing for hormone therapy for cases in which treatment was administered significantly early (<=60 days after diagnosis) for this measure had the highest concordance with 84.3 in 2006 and 79.1% for 2007 cases. There was 77.9% and 91.1% agreement in 2006 and 2007 diagnoses respectively for hormone therapy which was recommended but not administered for this measure. A total of 298 cases with missing hormone receptor status were reviewed, this information was found in nearly 90% of these cases.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

The NCDB collects all diagnosed cases within cancer programs. The measure exclusions as described in the specifications are the inverse of the measure inclusion criteria. Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure. These are established to ensure patients included in the

measure assessment meet the evidence based criteria. In 2012 85,570 breast cases were included in this measure.

The exception to this is the measure exclusion of, "Patient participation in a clinical trial which directly impacts the standard of care."

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:

An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed. For all cases applicable to this measure, in 2012 -2013, 5 cases were excluded from the measure denominator based on the exclusion of patient participation in a clinical trial which directly impacts the standard of care.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

Measure exclusions were used for n=5 (<0.01%) and does not affect estimated performance rates for this measure.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

This measure is specified to determine meaningful differences in compliance, differences in data performance was

described in performance gaps.

.

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:**

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**

This measure was not specified to report stratified performance rates, however the CoC´s recently released (2011) "real clinical time" Rapid Quality Reporting System (RQRS) (http://www.facs.org/cancer/ncdb/rqrs.html) reports back measure-specific performance rates by a number of strata, eg. patient age, sex, ethnicity, insurance status, and area-based SES. RQRS hosts a prospective treatment alert system, and so performance rates are both high and consistent with clinical expectation, however room for potential improvement remains. In a comparative analysis of 16 NCI/NCCCP pilot sites using RQRS with a comparative group of 25 other CoC-accredited cancer programs also using RQRS revealed that at NCCCP cancer programs white patients more frequently received HT (83.2%) than did African-American women (78.7%); and Medicaid recipients less frequently (76.2%) received HT than insured patients (83.3%). Comparative rates from the 25 non-NCCCP programs were slightly lower across the board, however the patterns of disparate receipt of care were mirrored

in this group of hospitals. Analysis from cases diagnosed 2008-2010. Disparities were reported in section 1.b of this application.
2.1-2.3 Supplemental Testing Methodology Information:
Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? **(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐** **Provide rationale based on specific subcriteria:**
If the Committee votes No, STOP

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)

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)*

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.**

The ACoS/CoC implementation of this measure is framed around the feasibility of data collection and reporting considerations. Cancer registries in the United States depend on a multitude of information sources in order to completely abstract case records and be in compliance with State, Federal and private sector accreditation requirements. Commission on Cancer Standards require case abstracting to be performed by a Certified Tumor Registrars (CTRs). CTRs must pass an exam and maintain continuing education. In the past decade, great strides have been made within the cancer registration community in terms of electronic capture of registry data from electronic pathology systems and electronic health records. However, until EHR systems are universally implemented in the US and fully integrated within hospital-level cancer registry systems, registry data will depend upon some level of human review and intervention to ensure data are complete and accurately recorded. Robust data quality edits are applied to the data at all levels of cancer data abstraction and processing. These edits standardize coded information and ensure its accuracy.

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.**

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

1) The infrastructure to monitor compliance with this measure has been in place since 2005 to assess and feed-back to approximately 1,500 Commission on Cancer (CoC) accredited centers performance rates for this measure. CoC accredited cancer programs account for 70-80% of patients affected by this measure. This measure is currently reported to CoC accredited programs through the National Cancer Data Base (NCDB) using the Cancer Program Practice Profile Report (CP3R) web-based audit and feed-back reporting tool. The CP3R is generally described at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. In addition, this measure is also reported to over 1000 cancer programs participating in its "real clinical time" feedback reporting tool through its Rapid Quality Response System (RQRS). An overview of the RQRS is available at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Both of these reporting tools have been utilized in the cancer registry community and do not produce an undue burden on the data collection network.

2) The data for this measure are key elements already collected in all hospital registries. This measure has been reviewed using cancer registry data. The CoC data demonstrates variation in the measure. The measure is readily implemented.

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)*.

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Pennsylvania Health Care Quality Alliance http://www.phcqa.org/ PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program https://www.medicare.gov/hospitalcompare/cancer-measures.html Regulatory and Accreditation Programs Commission on Cancer Accreditation https://www.facs.org/quality-programs/cancer/coc Professional Certification or Recognition Program QOPI® Certification Program http://www.instituteforquality.org/qcp/qopi-certification-measures Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Commission on Cancer, National Cancer Data Base https://www.facs.org/quality-programs/cancer/ncdb Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/manual-qopi-measures

4a.1. For each CURRENT use, checked above, provide:

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

Pennsylvania Health Care Quality Alliance
Purpose: The Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.

PPS-Exempt Cancer Hospital Quality Reporting program
Purpose: In 2010 the Affordable Care Act required the Centers for Medicare and Medicaid Services (CMS) to establish a specialized quality reporting program for the PPS-exempt cancer hospitals. The resulting PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program measures allow consumers to compare the quality of care given at the eleven PPS-exempt cancer hospitals currently participating in the program. This includes 11 PPS-Exempt Cancer Hospitals.

Commission on Cancer
Purpose: The Commission on Cancer (CoC) is a consortium of professional organizations dedicated to improving survival and quality of life for cancer patients through standard-setting, prevention, research, education, and the monitoring of comprehensive quality care. One of the standards for CoC-accredited cancer programs requires program meet established estimated performance rates

with accountability measures or develop an action plan for improvement. Standards Manual: https://www.facs.org/quality-programs/cancer/coc/standards. Approximately 1500 cancer programs are CoC-accredited constituting nearly 70% of diagnosed cases.

Commission on Cancer, National Cancer Data Base
Purpose: The National Cancer Data Base (NCDB) provides a benefit for CoC-accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP).

CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer and is described https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. This application is available to over 1500 CoC-accredited cancer programs

CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cqip

RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers - See more at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

Quality Oncology Practice Initiative
In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively

QOPI® Certification Program:
The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score. Please see a description of the QOPI® program above for details.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

EPRs, 95% CI, and number of patients by diagnosis year at the aggregate level are as follows:

2008 (80.0, 95% CI: 79.7-80.2, n=74,017)
2009 (83.9, 95% CI: 83.7-84.2, n=76.346)
2010 (88.5, 95% CI: 88.2-88.7, n=81,344)
2011 (89.5, 95% CI: 89.3-89.7, n=84,861)
2012 (87.8, 95% CI: 87.5-88.0, n=85,570)

There were 74,017 patients eligible for this measure in 2008 and 85,570 eligible in 2012. There were 1,487 facilities that had patients eligible for this measure in 2008, and 1,436 facilities with eligible patients in 2012. The EPR increased 7% between 2008 and 2012. In looking at EPRs at the hospital level, the mean EPR in 2008 was 78.7%, Standard Deviation 23.5, minimum=0, maximum=100%, interquartile range 72.3-94.4. The EPR in 2012 was 85.5, Standard deviation=19.8, minimum=0, maximum=100, interquartile range: 84.0-96.9.

EPRs increased in all census regions as described below.

Census Region. EPRs by census region were compared, for the Northeast, South, Midwest, West, and Pacific regions. EPRs increased in all regions between 2008 and 2012, with a 10% increase in the Northeast region, a 9% increase in the West and Pacific regions, a 7% increase in the South region, and a 6% increase in the Midwest region. In 2008, the lowest EPR was found in the Northeast (76.7, 95% CI: 76.0-77.4, n=15,568), followed by the Pacific region (77.5, 95% CI: 76.7-78.3, n=10,339), the South region (78.4, 95% CI: 77.9-79.0, n=25,754), the West region (79.2, 95% CI: 77.8-80.6, n=3,159), and the Midwest region (86.5, 95% CI: 86.0-87.0, n=19,025). In 2012, the lowest EPR was found in the South region (85.6, 85.2-86.0, n=30,981), followed by the Pacific region (86.5, 95% CI: 85.9-87.2, n=11,032), the Northeast region (86.9, 95% CI: 86.4-87.4, n=17,687), the West region (88.4, 95% CI: 87.5-89.4, n=4,300), and the Midwest region (92.2, 95% CI: 91.8-92.5, n=21,360).

4b.2. **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.**

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

This measure, as specified, is susceptible to under-reporting of the adjuvant hormone therapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. However, CoC accredited programs have demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. It does take additional time to collect and report this adjuvant therapy information. Additionally, the CoC´s Program Standards require review of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

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)**
0387 : Oncology: Hormonal therapy for stage IC through IIIC, ER/PR positive breast cancer

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

5a. Harmonization

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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
These measures are related but assess different levels of analysis and different data systems are used to determine eligibility and compliance.

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.)
0387 assesses hormone therapy for patients with stage Ic through III hormone receptor positive cancer. 0387 assesses if hormone therapy was prescribed within a 12 month period while our measure (0220) assesses if hormone therapy was administered within one year of diagnosis or if it was recommended but not received based on patient refusal, medical co-morbidity or other valid reasons.
0220 also assesses compliance at the facility level while 0387 assesses individual physician or practice level performance. The two measures use different data sources as well. 0220 utilizes cancer registry coding.

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):** Commission on Cancer, American College of Surgeons
Co.2 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-
Co.3 **Measure Developer if different from Measure Steward:** Commission on Cancer, American College of Surgeons
Co.4 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-

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 Developers: Christopher (Chris) Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); Richard Swanson, MD, FACS (Partners Health Care, Boston MA); Peter Enzinger, MD (Dana Farber Cancer Institute, Boston MA); Elin Sigurdson, MD, FACS (Fox Chase Cancer Center, Philadelphia PA); Mitchell Posner, MD, FACS (University of Chicago, Chicago IL); Anthony Robbins, MD, PhD (American Cancer Society)

The current Measure workgroup includes:
Charles Cheng MD, FACS (Fox Valley Surgical Associates, Appleton, WI), Daniel McKellar, MD, FACS (Wayne Healthcare, Greenville, OH), David Jason Bentrem, MD (Northwestern Memorial Hospital, Chicago, IL),
Karl Bilimoria, MD, FACS (Northwestern Univ/Feinberg Sch of Med, Chicago, IL), Lawrence Shulman MD (University of Pennsylvania, Philadelphia, PA), Matthew A Facktor, MD FACS (Geisinger Medical Center, Danville, PA), Ted James (University of Vermont, Burlington, VT)

This panel meets at least once annually to review quality measures currently supported and implemented by the ACoS Commission on Caner and to investigate and consider/review development of possible new measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 10, 2015
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?** 10, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 0223
Measure Title: Adjuvant chemotherapy is recommended or administered within 4 months (120 days) of diagnosis to patients under the age of 80 with AJCC III (lymph node positive) colon cancer
Measure Steward: Commission on Cancer, American College of Surgeons
Brief Description of Measure: Percentage of patients under the age of 80 with AJCC III (lymph node positive) colon cancer for whom adjuvant chemotherapy is recommended and not received or administered within 4 months (120 days) of diagnosis.
Developer Rationale: Improved survival for patients with Stage III lymph node positive colon cancer

Numerator Statement: Chemotherapy is administered within 4 months (120 days) of diagnosis or it is recommended and not received
Denominator Statement: Include, if all of the following characteristics are identified:
Age 18-79 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Primary tumors of the colon
Epithelial malignancy only
At least one pathologically examined regional lymph node positive for cancer (AJCC Stage III)
All or part of 1st course of treatment performed at the reporting facility
Known to be alive within 4 months (120 days) of diagnosis
Denominator Exclusions: Exclude, if any of the following characteristics are identified:
Age <18 and >=80; not a first or only cancer diagnosis; non-epithelial and non-invasive tumors; no regional lymph nodes pathologically examined; metastatic disease (AJCC Stage IV); not treated surgically; died within 4 months (120 days) of diagnosis; Patient participating in clinical trial which directly impacts receipt of standard of care.

Measure Type: Process
Data Source: Electronic Clinical Data : Registry, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 01, 2007 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement Preliminary Analysis

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 *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.

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:

- National Comprehensive Cancer Network (NCCN)Practice Guideline:
  - Pathologic Stage T1-3, N1-2, M0 or T4, N1-2, M0: FOLFOX or CapeOx (both category 1 and preferred). Other options include: FLOX (category 1) or Capecitabine or 5-FU/Leucovorin. **Level of evidence: Category 1** (defined as: based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate)
- Additional evidence included a systematic review of the body of evidence including multiple randomized clinical demonstrating approxate 25% reduction in risk of death.

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 updated the links for the guidelines and included the NCCN Categories of Evidence and Consensus – no changes were made to the evidence.

Exception to evidence
N/A

Guidance from the Evidence Algorithm
Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as high-level evidence (Box 6) → Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on 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.

- The developer provided the following national trend data from the National Cancer Data Base (NCDB):

	2008	2012
# of facilities	1,455	1,386
# of cases	12,207	10,517
Mean Performance Rate	82.0% (SD=0.23)	86.5% (SD=0.21)
IQR	72.1-100	80-100
Range	0-100	0-100

- The developer stated that more recent performance data was not provided because all adjuvant therapy information are likely incomplete for the most recent year until programs have had time to collect this information.
- In 2011, the Committee noted that the overall poor performance on this measure is concerning, given the very strong level 1 evidence of the impact on patient outcomes. A Committee member questioned whether Stage 2b colon cancers should be included in the measure. The developer explained the ability to identify that subset of Stage 2 colon cancers is not yet routinely possible due to the way the staging systems were designed until 2010, and stated the evidence is not settled regarding the appropriateness of adjuvant chemotherapy for Stage 2b disease.

Disparities:

Race/ethnicity

2012	Non-Hispanic white	Hispanic
# of cases	6,528	560
Mean performance rate	88.3%	77.7%

Age

2012	Age 18-49	Age 50-59	Age 60-69
# of cases	1,573	2,280	2,780
Mean performance rates	90.5%	87.1%	86.9%

- The developer provided additional disparities data on insurance status, income, facility type, and sex.

Questions for the Committee:

- *Without more recent performance data, does the Committee agree there is a gap in care that warrants a national performance measure?*

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Process measure that is a direct measure of the administration of adjuvant chemotherapy in patients with colon cancer likely to benefit from this treatment. A high level of evidence, Category 1 (defined as: based upon high-level evidence with is uniform NCCN consensus that the intervention is appropriate. Additional evidence included a systematic review of the body of evidence including multiple randomized clinical demonstrating approximate 25% reduction in risk of death. No new evidence presented. Algorithm rates this as MODERATE.
The evidence for this process measure directly relates to the process being measured. The Level I evidence related to the desired outcome of decreasing deaths due to colon cancer.

1b. Performance Gap
Comments:
No new performance data was provided with this submission. The developer stated that more recent performance data was not provided because all adjuvant therapy information is likely incomplete for the most recent year until programs have had time to collect this information. Of note, in 2011, the Committee felt that the overall poor performance on this measure was concerning, given the very strong level 1 evidence of the impact on patient outcomes. The performance gap in 2008 was 82.0% vs. 86.5% in 2012. Subgroup analysis was done with disparities in performance noted in most subgroups. Without new date, I rate this as MODERATE.
The performance data provided indicates a gap which is a less than optimal performance with a mean of less than 90% (86.5%). Disparities in performance are reported in race/ethnicity, age, insurance status, and income.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): registry and paper medical records

Specifications:

- This is a facility-level measure
- The numerator is defined as chemotherapy is administered within 4 months (120 days) of diagnosis or it is recommended and not received. Reasons chemotherapy is recommended and not received include:
    - Contraindicated due to patient risk factors
    - Patient died prior to planned or recommended therapy
    - Recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record.
    - Recommended by the patient´s physician, but treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record.
- The denominator includes:
    - Age 18-79 at time of diagnosis
    - Known or assumed to be first or only cancer diagnosis
    - Primary tumors of the colon
    - Epithelial malignancy only
    - At least one pathologically examined regional lymph node positive for cancer (AJCC Stage III)
    - All or part of 1st course of treatment performed at the reporting facility
    - Known to be alive within 4 months (120 days) of diagnosis
- Denominator exclusions include:
    - Age <18 and >=80
    - Not a first or only cancer diagnosis; non-epithelial and non-invasive tumors
    - No regional lymph nodes pathologically examined
    - Metastatic disease (AJCC Stage IV)
    - Not treated surgically; died within 4 months (120 days) of diagnosis
    - Patient participating in clinical trial which directly impacts receipt of standard of care - this is an update

since last endorsement date

- A calculation algorithm is provided.
- All cases which meet the measure criteria are included in the denominator. If a required data element is missing the case is flagged for additional review.
- Diagnosis codes are based on the Facility Oncology Registry Data Standards (FORDS), which were revised in 2016. Therefore, no ICD-9 or ICD-10 codes are provided for this measure.
- The database is a hospital cancer registry reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 1,400 cancer programs and approximately 65,200 cases from from all CoC-accredited cancer programs. The mean performance rates across all CoC-accredited cancer programs in 2007 was 88.3 and 88.1 in 2008. Cancer programs in the 75th percentile had performance rates of 100 in each respective year; 9.7% of programs had statistically low outlier performance rates (<58%), SD=21.3%.

Describe any updates to testing:

- The developer provided updated data from 2012:
  - 10,517 cases; 1,386 facilities
  - Mean performance rate (EPR): 86.5%
  - Standard Deviation: 0.21
  - IQR: 80-100
  - Minimum hospital-level performance rate: 0%
  - Maximum hospital-level performance rate: 100%
  - 189 programs reported EPRs in the lowest 10th percentile of 70.0%, with 419 programs in the lowest quartile with EPRs <= 85.7 %

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Overall performance rates do not meet criterion.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empirical testing as specified (Box 2) → empirical validity testing at patient level (Box 3)→use rating from validity testing of patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Questions for the Committee:

- *See questions under Validity*

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Randomly selected charts were reviewed by site surveyors to determine completeness and validity of data reported to registry. The measure denominator and numerator were viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Describe any updates to validity testing: The developer provided additional details on data element validity testing - see below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer provided the following information about the dataset: Survey sites and data collection occurred in 2009 and 2010. In 2009, 391 sites were reviewed and 5,712 charts – 322 of these charts were colon cases from 2006; representing 10.3% of measure eligible cases. In 2010, 423 sites were reviewed and 6,752 charts – 382 of these charts were colon cases from 2007; representing 10.6% of measure eligible cases.
- Data elements reviewed:
  - confirmation of timing of adjuvant therapy
  - documentation of treatment recommended but not received
  - assessment of missing and incomplete tumor characteristics

Validity testing results:

- The developer provided the following testing results:
  - "Assessment of timing for chemotherapy for cases in which treatment was significantly later than expected (>90 days after diagnosis) this measure had the highest concordance with 88.9% in 2006 diagnoses and 81.8% for 2007 cases. There was 88.5% and 92.4% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure. "

- The developer provided percentage agreement results for two of the data elements included in the numerator (timing of chemotherapy and therapy recommended but not received). NQF guidance states that testing should be done for all critical data elements.
- Site surveyors determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.
- Developers provided only percentage agreement statistics which indicated a decrease of 7.1% for the data element 'timing for chemotherapy' from 2006 (88.9%) to 2007 (81.8%); no additional results were provided (e.g., kappa scores, which indicate agreement over and above chance; sensitivity or specificity statistics).

Questions for the Committee:

- *Does the measure adequately identify and include colon cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exclude, if any of the following characteristics are identified:
  - Age <18 and >=80
  - Not a first or only cancer diagnosis; non-epithelial and non-invasive tumors
  - No regional lymph nodes pathologically examined
  - Metastatic disease (AJCC Stage IV)
  - Not treated surgically; died within 4 months (120 days) of diagnosis
  - Patient participating in clinical trial which directly impacts receipt of standard of care
- The measure exclusions as described are the opposite of the measure inclusion criteria. The cases excluded are those in which the clinical evidence does not support inclusion in the quality measure.
- In 2012-2013, 1 case (<0.01%) was excluded due to patient participating in clinical trial that directly impacts delivery of the standard of care; this exclusion does not affect estimated performance rates for this measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Performance data is presented above under opportunity for improvement. Complete details of the data are presented in 1b.

Question for the Committee:

- *Given the data provided in 1b., does the measure identify meaningful differences about quality across facilities?*

2b6. Comparability of data sources/methods:

- Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b7. Missing Data

- The developer describes in S.22 that all cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review. The developer does not provide information on the frequency of missing data or potential impact on results.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) → validity testing conducted with patient-level data elements (Box 10)→Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The data elements are clear with appropriate codes and logic for abstraction. I have no concerns about implementing this measure. There was some discussion at the last approval that the staging standards were in flux for Stage II colon cancer but this change has been noted since 2010 and should not interfere with measurement.

Data elements are clearly defined but reporting of percent agreement does not satisfy full conditions for assessing reliability. Overall performance rates do not meet criterion.

The specifications are consistent with the evidence.

Specifications are consistent with the evidence.

2a2. Reliability Testing

Comments:

The reliability algorithm suggests that the testing is insufficiently reliable and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff.

Scope of testing adequate. Percentage agreement provided for just two data elements not meeting NQF guidelines.

2b2. Validity Testing

Comments:

The validity algorithm suggests that the testing is insufficiently valid and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff. At this point, the measure does not pass the reliability and validity standards as set forth by NQF.

Scope of testing adequate.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Exclusions are appropriate and consistent with the evidence. No risk adjustment. This measure when implemented should be able to detect meaningful differences and should provide comparable results. Missing data is a threat. The developers discuss a strategy to identify and address missing data but no follow up is provided.

Exclusions are consistent with the evidence. No risk adjustment. Differences in performance indicate meaningful differences in quality of care for colon cancer.

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.

- Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry); some data elements are in defined fields in electronic sources.
- Data collection burden due to manual chart abstraction from paper medical records.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:

All of the data is routinely used during patient care and should be available in the medical record. The majority of the data requires abstraction which is fraught with potential area and burdensome to obtain. I have no concerns about the data collection strategy. I rate this MODERATE.

Data elements most likely generated during routine use of EHR although this may vary. Some burden exists if reliance upon CTRs for data generation.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details:

- PPS-Exempt Cancer Hospital Quality Reporting program: In 2010 the Affordable Care Act required the Centers for Medicare and Medicaid Services (CMS) to establish a specialized quality reporting program for the PPS-exempt cancer hospitals. The resulting PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program measures allow consumers to compare the quality of care given at the eleven PPS-exempt cancer hospitals currently participating in the program. This includes 11 PPS-Exempt Cancer Hospitals.
- Pennsylvania Health Care Quality Alliance (PHCQA): PHCQA is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. Develop a consensus-driven, statewide approach to hospital quality measurement that is supported by quality of care data from a variety of public data sources. We believe that by sharing aggregated quality performance data openly through public reporting on the Internet, we can provide valuable, objective health care quality information for all consumers. At the same time we can identify and share best practices to improve the performance of all stakeholders. Commission on Cancer (CoC) accredited cancer programs in Pennsylvania may elect to voluntarily report their estimated performance rates through this program. Currently 60 of 73 Pennsylvania programs are

participating.

- Quality Oncology Practice Initiative (QOPI®) (adapted): In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.
- Commission on Cancer, National Cancer Data Base: The National Cancer Data Base (NCDB) provides a venue for accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP). CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer. This application is available to over 1500 CoC-accredited cancer programs. CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

Improvement results:

- Overall, EPRs were higher in 2012 than in 2008:
  - 2008 81.7 (81.0 – 82.4)
  - 2009 84.9 (84.3 – 85.6)
  - 2010 88.2 (87.6 – 88.8)
  - 2011 89.3 (88.7 – 89.9)
  - 2012 86.5 (85.9 – 87.2)

Unexpected findings (positive or negative) during implementation:

- This measure, as specified, is susceptible to under-reporting of the adjuvant chemotherapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. Programs use of the CoC data quality tools has demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. Additionally, the CoC´s Program Standards require direct review and oversight of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

Potential harms:

- Developer did not identify any unintended consequences related to this measure.

Feedback :

- During public comment in 2012, commenters suggested that the measure could be improved upon by focusing only on administration of chemotherapy and not consideration of chemotherapy, as "considered" is not a precise term. The developer responded that the Commission on Cancer and the American College of Surgeons use cancer registries to implement this measure; the cancer registries have standard definitions for both "administered" and "considered" therapies. Cancer registries record and report this information if it is documented in the patient chart. Further, a review of data has demonstrated consistency in reporting considered therapies over three years.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

The measure is being publically reported and used in performance improvement activities. The importance of measuring the timely administration of adjuvant therapy for colon cancer will further the goal of high quality health care. I rate this HIGH.

Measure is publicly reported in Exempt Cancer Hospital Quality Reporting Program, the Pennsylvania Health Care Quality Alliance and the CoC National Cancer Data Base.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0385 : Oncology: Chemotherapy for AJCC Stage III Colon Cancer Patients

Harmonization

- 0385 assesses clinical group practice while 0223 assesses facility level performance.

Pre-meeting public and member comments

-

NQF #: 0223 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)***:**

Process

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Randomized trials that directly apply to this measure

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** Multiple randomized clinical trials

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* High level evidence

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* High level of consistency

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

Approximate 25% reduction in risk of death

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN)

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

1c.13 **Grade Assigned to the Body of Evidence:** Level I

1c.14 **Summary of Controversy/Contradictory Evidence:** None

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

See 1b.3

1c.16 **Quote verbatim, the specific guideline recommendation** *(Including guideline # and/or page #)***:**

Page COL 4

Pathologic Stage T1-3, N1-2, M0 or T4, N1-2, M0: FOLFOX or CapeOx (both category 1 and preferred). Other options include: FLOX (category 1) or Capecitabine or 5-FU/Leucovorin

1c.17 **Clinical Practice Guideline Citation:** National Comprehensive Cancer Network (NCCN)

1c.18 **National Guideline Clearinghouse or other URL:** http://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

1c.19 **Grading of Strength of Guideline Recommendation.** Has the recommendation been graded? Yes

1c.20 **If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN)

1c.21 **System Used for Grading the Strength of Guideline Recommendation:** Other

1c.22 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

1c.23 **Grade Assigned to the Recommendation:** Level I

1c.24 **Rationale for Using this Guideline Over Others:** All guidelines recommend chemotherapy with Stage III colon cancer

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0223_Evidence_MSF5.0_Data.doc,ACT_0223_Evidence_2016-635953597964456415.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Improved survival for patients with Stage III lymph node positive colon cancer

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The nationally recognized National Cancer Data Base (NCDB), jointly sponsored by the American College of Surgeons and the American Cancer Society, is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70 percent of newly diagnosed cancer cases nationwide and 30 million historical records. This analysis uses NCDB data.

The NCDB collects data from CoC accredited cancer programs on an annual basis; the data we collect is in accordance with standard registry procedures. In January of 2015, 2013 diagnoses were collected. This information was released to accredited cancer programs in the late summer. However, we find information on some of the therapies which take longer to be received are not complete upon initial submission and need time to document receipt of adjuvant therapy. Therefore the CoC does not begin surveying or holding programs accountable for their Estimated Performance Rates (EPRs) until the year after it is released to ensure

adequate adjuvant therapy information has been documented. We generally see a slight decrease in compliance for the most recent year until programs have had time to collect this information, since we don't feel all adjuvant therapy information are complete at initial submission we did not include the 2013 data in the application for this measure and used the next most recent annual rate of 2012 for this measure.

In 2008, 12,207 cases in 1,455 facilities were in the denominator and the mean estimated performance rate (EPR) was 82.0% (Std.=0.23). IQR=72.1-100, minimum=0, maximum=100. In 2012, 10,517 cases in 1,386 facilities were in the denominator and the mean estimated performance rate (EPR) was 86.5% (Std.=0.21). IQR=80-100, minimum=0, maximum=100.

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.
NA

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Data described in 1b.2
Disparities were assessed by race/ethnicity, age, insurance status, sex, facility type, and education and income at the zip code level.

Race/Ethnicity
Race/ethnicity was defined as non-Hispanic white, non-Hispanic black, Hispanic, Asian/Hawaiian/Pacific Island or other race/ethnicity. In all race/ethnicity groups, EPRs were higher in 2012 than in 2008. Non-Hispanic whites had the highest EPRs in both 2008 (EPR 83.5%, 95% CI 82.8-84.3, n=6954) and 2012 (88.3%, 95% CI 87.6-89.1, n=6528). Hispanics had the lowest EPR in both 2008 (73.9%, 95% CI 70.8-77.1, n=539) and 2012 (77.7%, 95% CI 74.6-80.7, n=560). Comparing 2012 to 2008, improvements in EPRs were as follows: 5% in Asian/Hawaiian/Pacific Islander, Non-Hispanic black, and Non-Hispanic white, 6% in 'Other' race, and 4% in Hispanic.

Age
Age groups were defined as, 18-49, 50-59, 60-69, 70-79. In all age groups, EPRs were higher in 2012 than in 2008. Age group 18-49 had the highest EPR in both 2008 (85.6%, 95% CI 84.0-87.2, n=1,860) and in 2012 (90.5%, 95% CI 89.1-91.9, n=1,573). As age increases, 2012 EPR decreases. Age groups 50-59 and 60-69 had the next highest EPRs at 87.1% (95% CI 85.8-88.4, n=2,280), and 86.9% (95% CI 85.7-88.0, n=2,780), respectively. Age group 70-79 had the lowest EPR both in 2008 (78.5%, 95% CI 77.1-79.8, n=3,662) and in 2012 (83.3%, 95% CI 82.0-84.7, n=2,469). Comparing 2012 to 2008, improvements in EPRs were as follows: 3% in ages 50-59, 5% in ages 18-49 and ages 70-79, and 6% in ages 60-69.

Insurance Status
Insurance status is defined as insurance at the time of diagnosis. Insurance was stratified into private, Medicare, Medicaid/ No insurance. In all insurance status groups, EPRs were higher in 2012 than in 2008. Private Insurance had the highest EPR in 2008 (84.7%, 95% CI 83.8-85.6, n=5998) and 2012 (88.5%, 95% CI 87.6-89.3, n=5,054). Comparing 2012 to 2008, the Not Insured/Medicaid category saw the largest improvement in EPR, at 7% (2008 EPR: 76.1%, 95% CI 73.2-79.0, n=821). During this time, improvements in EPR were 6% for Medicare and 4% for Private Insurance.

Median Income Quintile
Income quintiles at the zip code level were assessed based on the 2012 American Community Survey. In all median income quintiles, EPRs were higher in 2012 than in 2008. The highest EPR in 2012 was in the second highest income quintile, $53,000-68,999, (88.1%, 95% CI 86.8-89.4, n=2,459). The lowest income quintile, <$36,000, had the lowest EPR in 2012 (83.5%, 95% CI 81.7-85.3, n=1,628). Comparing 2012 with 2008, the lowest quintile (<$36,000) and the third quintile ($44,000-$52,999) saw the largest improvements in EPR at 6% (lowest quintile 2008 EPR: 77.4%, 95% CI 75.5-79.3, n=1,828), (third quintile 2008 EPR: 82.0%, 95% CI 80.4-83.5, n=2,387). The largest quintile ($69,000+) saw the smallest increase, at 3% (2008 EPR: 83.3%, 95% CI 81.9-84.7, n=2,761.

Facility Type
Facility type was assessed by programs CoC-accreditation status. Facility types include Comprehensive Community Cancer Programs, Integrated Network Cancer Programs, Community Cancer Programs and by Academic/Research programs. EPRs were higher in 2012 than in 2008 in every facility type. Comprehensive Community Cancer Programs had the highest EPRs in 2008 (84.0%,

95% CI 83.1-84.9, n=6,008) and in 2012 (88.0%, 95% CI 87.0-88.9, n=5,094). Community Cancer Programs had the lowest EPRs in 2012 (84.9%, 95% CI 83.1-86.7, n=1,521). Comparing 2012 with 2008, Academic/Research programs saw the greatest improvement in EPR, at 7% (2008 EPR: 77.8%, 95% CI 76.3-79.2, n=3,171). Community Cancer Programs, Comprehensive Community Cancer programs, and Integrated Network Cancer programs saw increases around 3-4%.

Sex
EPRs increased in both females and males from 2008 to 2013; 5% in females and 4% in males.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
NA

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Clear evidence that chemotherapy reduces the risk of distant disease recurrence and death in persons with node-positve (Stage III) colon cancer.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Gill S, Loprinzi CL, Sargent DJ et al. Pooled analyis of fluorouracil-based adjuvant therpay for Stage II and III Colon cancer;who benefits and by how much? J Clin Oncol 2004;22:1797-1806.
2. Sanoff HK, Carpenter WR, Martin CF, et al. Comparative effectiveness of oxaliplatin vs. Non-oxaliplatin-containing adjuvant chemtoherapy for stage III colon cancer. J Natl Cancer Inst. 2012;104:211-227.
3. JOnker DJ, Spithoff K, Maroun J. Adjuvant systemic chemotherapy for Stage II and III colon cancer after complete resection: an updated practice guideilne. Clin Oncol (R Coll Radiol) 2011;23:314-322.
4. Sharif S, O´Connell JF, Yothers G, Lopa S. Wolmark N. FOLFOX and FLOX regimens for the adjuvant treamtent of resected colon cancer. Cancer Invest. 2008;26:956-963.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Colorectal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Disparities

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.facs.org/~/media/files/quality%20programs/cancer/ncdb/measure%20specs%20colon_03312015.ashx

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Minor changes have been made to this measure. Cases are now excluded from the measure if participation in a clinical trial directly impacts the receipt of the standard of care and compliance with this measure. The title has been modified to state treatment is recommended rather than considered to be more consistent with the definitions in the registry codes used in specifying this measure.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Chemotherapy is administered within 4 months (120 days) of diagnosis or it is recommended and not received

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
4 months (120 days)

S.6. **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, 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.
Chemotherapy Recommended and not received [NAACCR Item#1390]=82-87 (82:not recommended/administered because it was contraindicated due to patient risk factors, 85:not administered because the patient died prior to planned or recommended therapy,86:It was recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record. 87: it was recommended by the patient´s physician, but this treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record)

OR; Chemotherapy [NAACCR Item#1390]=3, and Date Chemotherapy Started (NAACCR Item#1220] <=120 days following Date of Diagnosis [NAACCR Item# 340]

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Include, if all of the following characteristics are identified:
Age 18-79 at time of diagnosis
Known or assumed to be first or only cancer diagnosis
Primary tumors of the colon
Epithelial malignancy only
At least one pathologically examined regional lymph node positive for cancer (AJCC Stage III)
All or part of 1st course of treatment performed at the reporting facility

Known to be alive within 4 months (120 days) of diagnosis

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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 at Diagnosis [NAACCR Item#230] 18-79 AND Male or female [NAACCR item #220] = 1,2; AND Surgical Procedure of the Primary Site [NAACCR Item#1290] = 30–90, AND Regional Lymph Nodes Positive [NAACCR Item#820] = 1-90, 95, 97

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Exclude, if any of the following characteristics are identified:
Age <18 and >=80; not a first or only cancer diagnosis; non-epithelial and non-invasive tumors; no regional lymph nodes pathologically examined; metastatic disease (AJCC Stage IV); not treated surgically; died within 4 months (120 days) of diagnosis; Patient participating in clinical trial which directly impacts receipt of standard of care.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
See: https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/measure%20specs%20colon_03312015.ashx

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No stratification applied

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)

S.16. Type of score:
Rate/proportion
If other:

S.17. 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.18. Calculation Algorithm/Measure Logic *(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; aggregating data; risk adjustment; etc.)*
This measure score is calculated by dividing the numerator cases by denominator eligible cases.

Denominator eligible cases are assessed in a step-wise fashion:
- Include all colon cancer cases
- Adult patients 18 and over and under 80
- Males and female cases only
- Include first or only primaries
- Include epithelial tumors based on AJCC 7th Ed.
- Include invasive tumors only
- Exclude cases with clinical or pathologic evidence of in situ disease
- Exclude cases with clinical or pathologic evidence of metastatic disease
- Include only cases where all or part of first course treatment was performed at the reporting facility
- Include only surgically treated cases
- Include only patients which were alive for at least 120 days following diagnosis
- Include only lymph node positive disease

Numerator cases are then assessed from denominator eligible cases:
- Cases are included in the numerator if:
a) Chemotherapy is administered the number of days between diagnosis and start of chemotherapy within 120 days are included in the numerator or
b) Chemotherapy is recommended but not administered based on:
-Chemotherapy was not recommended/administered because it was contraindicated due to patient risk factors,
-Chemotherapy was not administered because the patient died prior to planned or recommended therapy,
-Chemotherapy was not administered. It was recommended by the patient´s physician but was not administered as part of the first course of therapy.
-Chemotherapy was not administered, it was recommended by the patients´ physician but refused by the patient, patient´s family member or guardian. The refusal was noted in patient record.

The measure score is calculated with the numerator divided by the denominator.

Detailed steps are found here:
https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/measure%20specs%20colon_03312015.ashx

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
All cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.

Hospital cancer registry data, reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0223_MeasureTesting_MSF5.0_Data.doc,0223_MeasureTesting_ACT_04012016-635953797306252412.doc

NATIONAL QUALITY FORUM

NQF #: 0223 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
This measure has been implemented by the ACoS CoC since 2007 across all CoC-accredited cancer programs, and reports on approximately 65,200 cases per year to almost 1,400 cancer programs.

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)*:
Cancer registry case records reported to the NCDB are reviewed annually, annualized hospital performance rates are provided back to CoC accredited cancer programs via the CoC´s Cancer Program Practice Profile Report (CP3R) using the denominator and numerator criteria documented in response to items 2a1.3 and 2a1.7, respectively, in the Specifications section. (http://www.facs.org/cancer/ncdb/cp3r.html)

2a2.3 **Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The mean performance rates across all CoC-accredited cancer programs was 88.3 in 2007 and 88.1 in 2008. The two years available at the time of this writing. Cancer programs in the 75th percentile had percentile had performance rates of 100 in each respective year. Even with high aggregate performance rates demonstrated by programs room for improvment across the system of CoC-accredited programs remains, with 9.7% of programs with statistically low outlier performance rates (<58%), SD=21.3%.

In 2012, 10,517 cases in 1,386 facilities were in the denominator and the mean estimated performance rate (EPR) was 86.5% (Std.=0.21). IQR=80-100, minimum=0, maximum=100. In 2012, 189 programs reported EPRs in the lowest 10th percentile of 70.0%, with 419 programs in the lowest quartile with EPRs <= 85.7 %

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
See 2a2.1. This measure has been implemented across all CoC-accredited cancer programs and subject to local review by standing committees of these hospitals and site surveyors at the time of accreditation site visits.

During Commission on Cancer Survey Site visits in 2009 and 2010, surveyors validated not more than 25 charts.
During 2009 – 391 accredited sites were reviewed, including 5,712 charts. This included an average of 14 charts per survey (IQR 6-22). 322 of these charts were colon cases; representing 10.3% of measure eligible cases.
During 2010- 423 accredited sites were reviewed, including 6,752 charts. This was based on an average of 14 charts per survey (IQR 6 – 22). 382 of these charts were colon cases; representing 10.6% of measure eligible cases.

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
Performance rates are reviewed and discussed, randomly selected charts are reviewed by the site surveyor to ascertain the completeness and validity of the data recorded in the local cancer registry and reported to the NCDB and included in the CP3R reporting application.

Major areas of review completed by site surveyors included but were not limited to, confirmation of timing of adjuvant therapy, documentation of treatment recommended but not received, assessment of missing and incomplete tumor characteristics.
2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
This measure has a high degree of user acceptability, the measure denominator and numerator are viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Assessment of timing for chemotherapy for cases in which treatment was significantly later than expected (>90 days after diagnosis) this measure had the highest concordance with 88.9% in 2006 diagnoses and 81.8% for 2007 cases. There was 88.5% and 92.4% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The NCDB collects all diagnosed cases within cancer programs. The measure exclusions as described in the specifications are the inverse of the measure inclusion criteria. Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure. These are established to ensure patients included in the measure assessment meet the evidence based criteria. In 2012, 10,517 colon cases were included in this measure.
The exception to this is the measure exclusion of, "Patient participation in a clinical trial which directly impacts the standard of care."
2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed. For all cases applicable to this measure, in 2012 -2013, 1 case was excluded from the measure denominator based on the exclusion of patient participation in a clinical trial which directly impacts the standard of care.
2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

Measure exclusions were used for n=1 (<0.01%) and does not affect estimated performance rates for this measure.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

This measure is specified to determine meaningful differences in compliance, differences in data performance was described in performance gaps.

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:**

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
This measure was not specified to report stratified performance rates, however the CoC´s recently released (2011) "real clinical time" Rapid Quality Reporting System (RQRS) (http://www.facs.org/cancer/ncdb/rqrs.html) reports back measure-specific performance rates by a number of strata, eg. patient age, sex, ethnicity, insurance status, and area-based SES. RQRS hosts a prospective treatment alert system, and so performance rates are both high and consistent with clinical expectation, however room for potential improvement remains. In a comparative analysis of 16 NCI/NCCCP pilot sites using RQRS with a comparative group of 25 other CoC-accredited cancer programs also using RQRS revealed that at NCCCP cancer programs female patients more frequently received adjuvant

chemotherapy (88.4%) than did males (84.9%); and interestingly patients living in income-disadvantaged areas more frequently (89.9%) received adjuvant chemotherapy than did patients residing in more affluent areas (85.4%). Comparative rates from the 25 non-NCCCP programs were slightly less pronounced between patient sex and income SES. Analysis from cases diagnosed 2008-2010.

Disparities were reported in section 1.b of this application.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? (*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

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)
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)*
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.**
The ACoS/CoC implementation of this measure is framed around the feasibility of data collection and reporting considerations. Cancer registries in the United States depend on a multitude of information sources in order to completely abstract case records and be in compliance with State, Federal and private sector accreditation requirements. Commission on Cancer Standards require case abstracting to be performed by a Certified Tumor Registrars (CTRs). CTRs must pass an exam and maintain continuing education. In the past decade, great strides have been made within the cancer registration community in terms of electronic capture of registry data from electronic pathology systems and electronic health records. However, until EHR systems are universally implemented in the US and fully integrated within hospital-level cancer registry systems, registry data will depend upon some level of human review and intervention to ensure data are complete and accurately recorded. Robust data quality edits are applied to the data at all levels of cancer data abstraction and processing. These edits standardize coded information and ensure its accuracy.

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.**
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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The infrastructure to monitor compliance with this measure has been in place since 2005 to assess and feed-back to approximately 1,500 Commission on Cancer (CoC) accredited centers performance rates for this measure. CoC accredited cancer programs account for 70-80% of patients affected by this measure. This measure is currently reported to CoC accredited programs through the National Cancer Data Base (NCDB) using the Cancer Program Practice Profile Report (CP3R) web-based audit and feed-back reporting tool. The CP3R is generally described at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. In addition, this measure is also reported to over 1030 cancer programs participating in its "real clinical time" feedback reporting tool through its Rapid Quality Response System (RQRS). An overview of the RQRS is available at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Both of these reporting tools have been utilized in the cancer registry community and do not produce an undue burden on the data collection network. Utilization of these tools increases the completeness of adjuvant therapy information captured by the cancer registry, cancer registries need time to obtain this complete adjuvant therapy information.

The data for this measure are key elements already collected in all hospital registries. This measure has been reviewed using cancer registry data. The CoC data demonstrates variation in the measure. Registries have demonstrated the ability to identify gaps in data collection and to correctly identify therapy in the majority of cases. The measure is readily implemented.

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)*.

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.

Planned	Current Use (for current use provide URL)
Regulatory and Accreditation Programs Quality Improvement (Internal to the specific organization)	Public Reporting PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program Pennsylvania Health Care Quality Alliance https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228774283195 Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/quality-oncology-practice-initiative-qopi Commission on Cancer, National Cancer Data Base

	https://www.facs.org/quality-programs/cancer/ncdb

4a.1. **For each CURRENT use, checked above, provide:**

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

4.1.a Public Reporting

PPS-Exempt Cancer Hospital Quality Reporting program
Purpose: In 2010 the Affordable Care Act required the Centers for Medicare and Medicaid Services (CMS) to establish a specialized quality reporting program for the PPS-exempt cancer hospitals. The resulting PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) Program measures allow consumers to compare the quality of care given at the eleven PPS-exempt cancer hospitals currently participating in the program. This includes 11 PPS-Exempt Cancer Hospitals.

Pennsylvania Health Care Quality Alliance (PHCQA)
About: PHCQA is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. Develop a consensus-driven, statewide approach to hospital quality measurement that is supported by quality of care data from a variety of public data sources. We believe that by sharing aggregated quality performance data openly through public reporting on the Internet, we can provide valuable, objective health care quality information for all consumers. At the same time we can identify and share best practices to improve the performance of all stakeholders. Commission on Cancer (CoC) accredited cancer programs in Pennsylvania may elect to voluntarily report their estimated performance rates through this program. Currently 60 of 73 Pennsylvania programs are participating.

4.1.f Quality Improvement with Benchmarking

Quality Oncology Practice Initiative (QOPI®) (adapted)
In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

Commission on Cancer, National Cancer Data Base
Purpose: The National Cancer Data Base (NCDB) provides a venue for accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP).

CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer and is described https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3rv. This application is available to over 1500 CoC-accredited cancer programs

CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cqip

RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers - See more at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

4a.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?*)
NA

4a.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.*)

NA

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Progress

2008 81.7 (81.0 – 82.4)
2009 84.9 (84.3 – 85.6)
2010 88.2 (87.6 – 88.8)
2011 89.3 (88.7 – 89.9)
2012 86.5 (85.9 – 87.2)

Overall, EPRs were higher in 2012 (mean EPR 86.5%, 12,207 cases, 1,455 facilities) than in 2008 (mean EPR 82%, 10,517 cases, 1,386 facilities): a 4.5% increase. Comparing 2012 to 2008, every race/ethnicity, age group, facility type, insurance status group, and income and education group saw an improvement in EPR; some groups saw larger increases in EPR than others, and there was some variation in EPR among groups. In 2012, Non-Hispanic white had the highest EPR (88.3%, 95% CI 87.6-89.1, n=6528), and Hispanic had the lowest (77.7%, 95% CI 74.6-80.7, n=560), and all race/ethnicity groups saw an improvement in EPR around 4-6%. Ages 18-49 had the highest EPR in 2012 (90.5%, 95% CI 89.1-91.9, n=1,573), and EPR decreased as age increased, to the lowest; the EPR in ages 70-79 was 83.3% (95% CI 82.0-84.7, n=2,469). EPRs increased 3-6% among all age groups from 2008 to 2012. 2012 EPRs for insurance status groups ranged from 83% to 89%. Private Insurance had the highest EPR in 2012 (88.5%, 95% CI 87.6-89.3, n=5,054). Not Insured/ Medicaid saw the highest improvement (7%) and had the lowest EPR (2012 EPR: 83.4%, 95% CI: 81.0-85.9, n=880). EPRs for facility types in 2012 ranged from 85% to 88%; the highest being Comprehensive Community Cancer Programs (88.0%, 95% CI 87.0-88.9, n=5,094) and the lowest being Academic/Research and Community Cancer Programs. All types saw increases from 2008-2012 around 3-4%, except Academic/Research, which saw a 7% increase (2012 EPR: 84.9%, 95% CI 83.7-86.2, n=3,027). Aside from the lowest quintile, 2012 EPRs for income were 86-88%; the EPR for <$36,000 was 84%. The largest income group ($69,000+) saw the highest 2012 EPR and the smallest increases in EPRs from 2008 to 2012 (3%), and the remaining income and education groups saw increases around 4-7%.

Geographic Area: Census Division

EPRs were higher in 2012 than in 2008 in every Census Division. In 2012, the highest EPRs were in West North Central (92.9%, 95% CI 91.1-94.8, n=750) and New England (92.5%, 95% CI 90.2-94.8, n=495). The next highest EPRs were at 90%: East North Central and East South Central, followed by South Atlantic, Mountain, Middle Atlantic, Pacific, which had EPRs around 83-86%. West South Central had the lowest EPR (79.4%, 95% CI 77.0-81.9, n=1,065). The divisions that saw the largest improvements in EPR in 2012 compared to 2008 were Mountain (2008 EPR: 76.7%, 95% CI 71.2-80.2, n=558), Middle Atlantic (2008 EPR: 76.7%, 95% CI 74.7-78.7, n=1,720), and East South Central (2008 EPR: 81.9%, 95% CI 79.3-84.5, n=850), each with an improvement by 8-9%. Pacific, East North Central, West South Central, South Atlantic, and West North Central saw the smallest increases, at 3-4%. New England saw an improvement of 6%.

4b.2. **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.**

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
This measure, as specified, is susceptible to under-reporting of the adjuvant chemotherapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. Programs use of the CoC data quality tools has demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. Additionally, the CoC´s Program Standards require direct review and oversight of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

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)**
0385 : Oncology: Chemotherapy for AJCC Stage III Colon Cancer Patients

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

5a. Harmonization
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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The measures assess different levels of data analysis, 0385 assesses clinical group practice while 0223 assesses facility level performance. The data sources are also different for the two measures increasing the burden of collection for harmonization.

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 target populations of these measures and the level of analysis are sufficiently different to warrant both measures. Measure 0223 assesses adjuvant chemotherapy on surgically treated patients to be reported at the facility level for CoC-accredited cancer

programs.
Measure 0223 assesses receipt of chemotherapy based on information captured through cancer registries utilizing coding of the North American Association of Central Cancer Registries (NAACCR) while measure 0385 assesses compliance utilizing CPT codes through clinical practices.

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):** Commission on Cancer, American College of Surgeons
Co.2 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-
Co.3 **Measure Developer if different from Measure Steward:** Commission on Cancer, American College of Surgeons
Co.4 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-

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 developers:
Christopher (Chris) Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); Richard Swanson, MD, FACS (Partners Health Care, Boston MA); Peter Enzinger, MD (Dana Farber Cancer Institute, Boston MA); Elin Sigurdson, MD, FACS (Fox Chase Cancer Center, Philadelphia PA); Mitchell Posner, MD, FACS (University of Chicago, Chicago IL); Anthony Robbins, MD, PhD (American Cancer Society)

The current Measure workgroup includes:
Charles Cheng MD, FACS (Fox Valley Surgical Associates, Appleton, WI), Daniel McKellar, MD, FACS (Wayne Healthcare, Greenville, OH), David Jason Bentrem, MD (Northwestern Memorial Hospital, Chicago, IL), Karl Bilimoria, MD, FACS (Northwestern Univ/Feinberg Sch of Med, Chicago, IL), Lawrence Shulman MD (University of Pennsylvania, Philadelphia, PA), Matthew A Facktor, MD FACS (Geisinger Medical Center, Danville, PA), Ted James (University of Vermont, Burlington, VT)

This panel meets at least once annually to review quality measures currently supported and implemented by the ACoS Commission on Caner and to investigate and consider/review development of possible new measures.Christopher (Chris) Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); Richard Swanson, MD, FACS (Partners Health Care, Boston MA); Peter Enzinger, MD (Dana Farber Cancer Institute, Boston MA); Elin Sigurdson, MD, FACS (Fox Chase Cancer Center, Philadelphia PA); Mitchell Posner, MD, FACS (University of Chicago, Chicago IL); Anthony Robbins, MD, PhD (American Cancer Society)

This panel meets at least once a calendar quarter to review quality measures currently supported and implemented by the ACoS Commission on Caner and to invstigate and consider/review development of possible new measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 10, 2015
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?** 10, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 0225
Measure Title: At least 12 regional lymph nodes are removed and pathologically examined for resected colon cancer.
Measure Steward: Commission on Cancer, American College of Surgeons
Brief Description of Measure: Percentage of patients >18yrs of age, who have primary colon tumors (epithelial malignancies only), at AJCC stage I, II or III who have at least 12 regional lymph nodes removed and pathologically examined for resected colon cancer.
Developer Rationale: Improved survival for patients with a greater number of lymph nodes resected; greater accuracy of staging for patients, and consequently appropriate post-surgical care

Numerator Statement: >=12 regional lymph nodes pathologically examined.
Denominator Statement: Include, if all of the following characteristics are identified:
Age >=18 at time of diagnosis
Primary tumors of the colon
Epithelial malignancy only
AJCC Stage I, II, or III
Surgical resection performed at the reporting facility
Denominator Exclusions: Exclude, if any of the following characteristics are identified:
Age <18; non-epithelial and non-invasive tumors; metastatic disease (AJCC Stage IV); not treated surgically at the reporting facility; perforation of the primary site; acute obstruction

Measure Type: Process
Data Source: Electronic Clinical Data : Registry, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 01, 2007 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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:

- National Comprehensive Cancer Network (NCCN) Practice Guideline:
  - For stage II (pN0) colon cancer, if less than 12 lymph nodes are initially identified, it is recommended that the pathologist go back to the specimen and resubmit more tissue of potential lymph nodes. If 12 lymph nodes are still not identified, a comment in the report should indicate that an extensive search for lymph nodes was undertaken. **Level of evidence: Category: 2A** (Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.)
- Additional evidence included a systematic review of the body of evidence including multiple observational studies. A summary of the evidence concluded that:
  - There is a lack of consensus as to the minimal number of lymph nodes that necessarily have to be examined to accurately identify AJCC stage III colon cancer; and
  - Studies using registry/administrative data have shown that the proportion of patients within a hospital who undergo an "adequate" lymph node examination may not be associated with a survival benefit at the hospital level.
- In 2012, the Steering Committee noted that lower level quality of evidence was presented. A large body of observational studies was provided in support of the measure, but no RCTs. The Steering Committee was concerned that some literature suggests that removal of anywhere from 6 to 17 nodes is the appropriate number. The developer noted that was true; however, NCCN guidelines call for 12 lymph nodes. The developer noted that this will be a moving target, and as the literature on the topic improves, the measure will be updated accordingly.

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 updated the links for the guidelines and included the NCCN Categories of Evidence and Consensus – no changes were made to the evidence.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as lower-level evidence/observational studies only (Box 6) → Low (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review.*
- *Is the SC aware of higher-level evidence to support this measure?*
- *Does the SC think there is a need to repeat the discussion and voting on 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.

- The developer provided the following national trend data from the National Cancer Data Base (NCDB):

	2008	2013	2008-2013
# of cases	39,910	41,546	--
Mean Performance Rate	81.7%	89.7%	--
IQR	71-92%	84-97%	--
Min,Max	--	--	0-100%

Disparities:

Race/ethnicity

2013	Non-Hispanic white	Non-Hispanic black	Hispanic	Asian/Hawaiian/Pacific Island
# of cases	31,698	4,752	2,177	1,064
Performance Rates	89.9%	88.6%	89.7%	90.6%

Age

2013	Age 18-49	Age 50-59	Age 60-69
# of cases	3,164	6,776	9,762
Performance Rates	94.5%	90.0%	89.7%

- The developer provided additional disparities data on insurance status, income, high school degree, facility type, and census type.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The evidence for this process indicator relates directly but is of lower rating. Systematic review included but many observational studies included and there is not final consensus on number of nodes. Related to desired outcomes as it is believed it will lead to improved identification of colon cancer.

Process measure. Tangentially related to the outcome since there is less evidence (small studies, retrospective studies) to support the link between survival and lymph node dissection. And, the cause and effect is not clear. Is it a surgical problem or a pathology evaluation problem? The level of evidence is Category 2A (Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.) A large body of observational studies was provided in support of the measure, but no RCTs. In 2012, the developer noted that this will be a moving target, and as the literature on the topic improves, the measure will be updated accordingly. However, no new data was presented. Numerous studies and sufficient data support that retrieval of more lymph nodes is associated with improved survival for individual patients, but the causality underlying this relationship is unknown. It is important to adhere to strict oncologic principles for cancer resections, including high vascular ligation and complete en bloc resection of the meso-colon, lymphadenectomy, and circumferential margins (for rectal cancer). In addition, it is important for pathologists to perform diligent examination of resected specimens. I grade the evidence as LOW. This is confirmed by the algorithm.

1b. Performance Gap

Comments:

There is a moderate opportunity for improvement with disparities in age noted. Overall performance rate is 89.7% increased from 81.7%.

New data in the form of national trending data was provided. National trending data show improvement from the first approval to a report from 2013 from 81.7%-89.7%. The range remains from 0-100% from 2008-2013. No definitive evidence that this data is topped off because of the improvement in the numbers but it is unlikely that the percentage will increase much in high performers since medical variation is likely to account for a significant number of patients that don't get to the 12 node goal. There are definitely low performing hospitals in this trending data. Limited evidence for a disparity gap was provided. I rate this as MODERATE.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): registry and paper medical records

Specifications:

- This is a facility-level measure
- The developer described several updates to the specifications
- The numerator is defined as >=12 regional lymph nodes pathologically examined
- The denominator is defined as:
  - Age >=18 at time of diagnosis
  - Primary tumors of the colon
  - Epithelial malignancy only
  - AJCC Stage I, II, or III
  - Surgical resection performed at the reporting facility
- Denominator exclusions include:
  - Age <18
  - Non-epithelial and non-invasive tumors
  - Metastatic disease (ajcc stage iv)
  - Not treated surgically at the reporting facility
  - Perforation of the primary site
  - Acute obstruction
- A calculation algorithm is provided.
- All cases which meet the measure criteria are included in the denominator. If a required data element is missing the case is flagged for additional review.
- Diagnosis codes are based on the Facility Oncology Registry Data Standards (FORDS), which were revised in 2016. Therefore, no ICD-9 or ICD-10 codes are provided for this measure.
- The database is a hospital cancer registry reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing attachment

Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 1,400 cancer programs and approximately 37,800 cases from from all CoC-accredited cancer programs. The median performance rate across all CoC-accredited cancer programs in 2007 was almost 79.0 and the mean performance performance rate was 75.0. In 2008 the mean performance rate was 80.4 and the medican was 83.3. The mean increased to 81.5 in 2009 with a median of 85.7. In 2008, 2.5% (n=34) of programs had a performance rate below 41.0%.

Describe any updates to testing:

- The developer provided updated data from 2013:
  - # of cases: 41,546
  - Mean performance rate: 89.7% (89.4-90.0)
  - IQR: 84.0-97.0%
  - 2008-2013 min, max hospital-level performance rate: 0-100%
  - 189 programs in 2013 reported compliance <=75.0%, representing the lowest 10th percentile of programs.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Overall performance rates do not meet criterion.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empirical testing as specified (Box 2) → empirical validity testing at patient level (Box 3)→use rating from validity testing of patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Questions for the Committee:

- *See questions under Validity*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Randomly selected charts were reviewed by site surveyors to determine completeness and validity of data reported to registry. The measure denominator and numerator were viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Describe any updates to validity testing: The developer provided additional details on data element validity testing - see below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**

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

Validity testing method:

- The developer provided the following information about the dataset: Survey sites and data collection occurred in 2009 and 2010. In 2009, 391 sites were reviewed and 5,712 charts – 322 of these charts were colon cases from 2006; representing 10.3% of measure eligible cases. In 2010, 423 sites were reviewed and 6,752 charts from 2007 – 382 of these charts were colon cases; representing 10.6% of measure eligible cases.
- Data elements reviewed:
  - confirmation of timing of adjuvant therapy
  - documentation of treatment recommended but not received
  - assessment of missing and incomplete tumor characteristics

Validity testing results:

- The developer provided the following testing results:
  - "Assessment of timing for chemotherapy for cases in which there were positive lymph nodes found where treatment was significantly later than expected (>90 days after diagnosis) this measure had the highest concordance with 88.9% in 2006 diagnoses and 81.8% for 2007 cases. There was 88.5% and 92.4% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure. "
- The developer provided percentage agreement results for two of the data elements included in the numerator (timing of chemotherapy and therapy recommended but not received). NQF guidance states that testing should be done for all critical data elements.
- Site surveyors determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.
- Developers provided only percentage agreement statistics which indicated a decrease of 7.1% for the data element 'timing for chemotherapy' from 2006 (88.9%) to 2007 (81.8%); no additional results were provided (e.g., kappa scores, which indicate agreement over and above chance; sensitivity or specificity statistics).

Questions for the Committee:

- *Does the measure adequately identify and include colon cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exclude, if any of the following characteristics are identified:
  - Age <18

- Non-epithelial and non-invasive tumors
- Metastatic disease (ajcc stage iv)
- Not treated surgically at the reporting facility
- Perforation of the primary site
- Acute obstruction

- Performation of the primary site and acute obstruction have been added to account for emergent cases.
- The measure exclusions as described are the opposite of the measure inclusion criteria. The cases excluded are those in which the clinical evidence does not support inclusion in the quality measure.
- The developer stated, "An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed." though, this is not an exclusion for this measure.
- Additionally, the developer stated that in 2012 -2013, 16 cases were excluded for Perforation of the primary site and 26 cases were excluded based on acute obstructions. This is a total of 42 of 41,546 (<0.01%) eligible cases excluded from this measure; this exclusion does not affect estimated performance rates for this measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Performance data is presented above under opportunity for improvement. Complete details of the data are presented in 1b.

Question for the Committee:

- *Given the data provided in 1b., does the measure identify meaningful differences about quality across facilities?*

2b6. Comparability of data sources/methods:

- Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b7. Missing Data

- The developer describes in S.22 that all cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review. The developer does not provide information on the frequency of missing data or potential impact on results.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) → validity testing conducted with patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Preliminary rating from validity algorithm: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Data elements are clearly defined and the measure can be consistently implemented but overall reliability rating is insufficient as percent agreement provided only for 2 data elements.

The specifications clearly define the appropriate data elements, codes, and calculations. I have no concerns about the ability to implement this measure.
The specifications are consistent with the evidence.

2a2. Reliability Testing
Comments:
Adequate scope but overall reliability rating is insufficient as percent agreement provided only for 2 data elements.
The reliability algorithm suggests that the testing is insufficiently reliable and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff.

2b2. Validity Testing
Comments:
Adequate scope. I agree that this measure reflects an aspect of quality despite difference of opinion on number of nodes collected, there is wide acceptance of this measure given current state of the evidence.
The validity algorithm suggests that the testing is insufficiently valid and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff. At this point, the measure does not pass the reliability and validity standards as set forth by NQF.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Exclusions are consistent with the evidence. No risk adjustment. Missing data presents minimal threat to validity as missing data elements are flagged for review.
The exclusions are appropriate to define the correct population. There is no risk adjustment. It appears that the analyses can demonstrate meaningful differences since there is a standard methodology across all sites and a credentialing process for the sites. The analyses should allow for comparisons. However, in all cases, there is a plan for missing data but further information regarding the success of acquiring this data is not provided with this submission.

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.

- Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry); some data elements are in defined fields in electronic sources.
- Data collection burden due to manual chart abstraction from paper medical records.

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

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:

Most data elements likely routinely recorded in EHRs. Reliance on CTRs presents some level of burden for organizations. Minimal concern over data collection strategy given quickly evolving use of EHRs.
All of the data is routinely used during patient care and should be available in the medical record. The majority of the data requires abstraction which is fraught with potential area and burdensome to obtain. The other feasibility issue is that of creating a collaborative atmosphere between the surgeon and the pathologist since the greatest gains in increasing the number of lymph nodes dissected can sometime be by increasing the quality of the pathologic dissection. I have no concerns about the data collection strategy. I rate this MODERATE.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- Pennsylvania Health Care Quality Alliance (PHCQA): PHCQA is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. This organization works to develop a consensus-driven, statewide approach to hospital quality measurement that is supported by quality of care data from a variety of public data sources. Commission on Cancer (CoC) accredited cancer programs in Pennsylvania may elect to voluntarily report their estimated performance rates through this program. Currently 60 of 73 Pennsylvania programs are participating.
- Commission on Cancer Accreditation: The CoC standards have a requirement for programs to meet expected estimated performance rates for accountability measures approved through the CoC (Standard 4.5), programs are required to meet 85% compliance with this measure, either through the estimated performance rate point estimate or the 95% confidence interval. If this benchmark was not met then programs were required to develop action plans to address this issue. There are over 1500 CoC-accredited cancer programs representing approximately 70% of newly diagnosed cancer cases annually.
- Commission on Cancer, National Cancer Data Base: The National Cancer Data Base (NCDB) provides a venue for accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP). CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer. This application is available to over 1500 CoC-accredited cancer programs. CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.
- Quality Oncology Practice Initiative : In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

Improvement results:

- 2008: 81.8 (81.4 – 82.2) N = 39,910
- 2009: 84.6 (84.3 – 85.0) N = 38,578
- 2010: 86.3 (85.9 – 86.6) N = 36,145
- 2011: 87.6 (87.3 – 88.0) N= 35,738
- 2012: 88.1 (87.8 – 88.4) N= 42,570
- 2013: 89.7 (89.4 – 90.0) N= 41,546

Unexpected findings (positive or negative) during implementation :

- This measure, as specified, is unlikely to be systematically susceptible to under-reporting due to the integral dependence of the measure upon information routinely documented and reported following pathologic examination of colon tissue specimens.

Potential harms:

- Developer did not identify any unintended consequences related to this measure.

Feedback :

- No additional feedback received.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Publicly reported in Pennsylvania Health Care Quality Alliance and Commission on Cancer NDB.
This measure is being used for public reporting and quality improvement. The potential value is that of creating a collaborative atmosphere between the surgeon and the pathologist since the greatest gains in increasing the number of lymph nodes dissected can sometimes be by increasing the quality of the pathologic dissection. I rate this as MODERATE.

Criterion 5: Related and Competing Measures

Related or competing measures

- None identified

Harmonization

Pre-meeting public and member comments

-

NQF #: 0225 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)***:**

Process

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Observational studies

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** Multiple observational studies

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* Medium/High level evidence

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Moderate to high level of consistency

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

1c.9 **Grading of Strength/Quality of the Body of Evidence.** Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN)

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Level I, IIA, IIB, III

NCCN Categories of Evidence and Consensus
Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.
Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.
Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.
Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.
All recommendations are category 2A unless otherwise noted.
http://www.nccn.org/professionals/physician_gls/categories_of_consensus.asp

1c.13 **Grade Assigned to the Body of Evidence:** IIA

1c.14 **Summary of Controversy/Contradictory Evidence:** 1. There is a lack of consensus as to the minimal number of lymph nodes

that necessarily have to be examined to acurately identify AJCC stage III colon cancer. 2. Studies using registry/administrative data have shown that the proportion of patients within a hospital who undergo an "adequate" lymph node examination may not be associated with a survival benefit at the hospital level.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

See 1b.3

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

COL-A-3 of 5

For stage II (pN0) colon cancer, if less than 12 lymph nodes are initially identified, it is recommended that the pathologist go back to the specimen and resubmit more tissue of potential lymph nodes. If 12 lymph nodes are still not identified, a comment in the report should indicate that an extensive search for lymph nodes was undertaken.

1c.17 Clinical Practice Guideline Citation: NCCN Clinical Practice Guidelines - www.nccn.org

1c.18 National Guideline Clearinghouse or other URL: http://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: National Comprehensive Cancer Network (NCCN)

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: Level I, IIA, IIB, III

1c.23 Grade Assigned to the Recommendation: IIA

1c.24 Rationale for Using this Guideline Over Others: Broad recognition of the NCCN clinical guidelines as the "gold-standard".

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: Moderate**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0225_Evidence_MSF5.0_Data.doc,12RLN_0225_Evidence_2016_04052016.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Improved survival for patients with a greater number of lymph nodes resected; greater accuracy of staging for patients, and consequently appropriate post-surgical care

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The nationally recognized National Cancer Data Base (NCDB), jointly sponsored by the American College of Surgeons and the American Cancer Society, is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70 percent of newly diagnosed cancer cases nationwide and 33 million historical records. Data from the NCDB was analyzed.

The NCDB collects data from CoC accredited cancer programs on an annual basis; the data we collect is in accordance with standard registry procedures. In January of 2015, 2013 diagnoses were collected. This information was released to accredited cancer programs in the late summer 2015, this data is used in this application.

The mean performance rate for this measure has increased from 81.7% (95% CI: 81.4-82.1) IQR=71-92% n= 39,910 in 2008 to 89.7% (89.4-90.0) IQR=84-97% n=41,546 in 2013 representing a steady improvement in quality. The minimum hospital-level performance rate is 0% with a 100% maximum in all years assessed 2008-2013.

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.
NA

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
The data source is described in 1b.1.
Disparities were assessed by race/ethnicity, age, insurance status, sex, facility type, andincome at the zip code level.

Race/Ethnicity
Race/ethnicity was defined as non-Hispanic white, non-Hispanic black, Hispanic, Asian/Hawaiian/Pacific Island or other race/ethnicity. Between 2008 and 2012, performance rates increased in all ethnic groups. Non-hispanic whites, Non-hispanic blacks, Asian Pacific Islanders, Hispanics, and Non-hispanic Blacks had similar performance rates in 2013; Non-hispanic whites at 89.9% (95% CI: 89.6-90.3) n=31,698 in 2013, non-Hispanic Blacks at 88.6% (95% CI: 87.7-89.6) n=4752, Asian Pacific Islanders 90.6% (88.9-92.4) n=1064, and Hispanics 89.7% (95% CI: 88.4-90.9) n=2177.

Age
Age groups were defined as, 18-49, 50-59, 60-69, 70-79 and over 79. Since 2008, patient above age 49 saw a relatively equal gain in performance with the measure ~9%. Patients age 18-49 saw only a 5.2% increase given the already high performance 89.3% (88.3-90.4) n=3360 in 2008. Patients under the age of 50 at diagnosis had higher performance rates in 2013 at 94.5% (93.7-95.2) n=3164, compared to patients aged 50 to 59 90.0% (89.3-90.7) n=6776, patients aged 60-69 89.7% (89.1-90.3) n=9762, patients aged 70-79 89.2% (88.7-89.8) n=11,196, and patients aged >79 88.6% (88.0-89.2) n=10,648.

Insurance
Insurance status is defined as insurance at the time of diagnosis. Insurance was stratified into private, Medicare, Medicaid/No insurance. Since 2008, patients with each insurance type saw a relative equal gain in performance of 8-10%. Uninsured and insured including Medicaid patients had similar performance rates in 2013. Uninsured at 91.2% (90.1-92.4) n=2310, private insurance at 90.7% (90.2-91.2) n=13,461, Medicare at 89.0% (88.6-89.4) n=24,792, Other Government 89.7% (86.6-92.9).

Median Zip Code Income Quintile
Income quintiles at the zip code level were assessed based on the 2012 American Community Survey. Patients that resided in communities with a median income of <$36,000 annually at diagnosis experienced higher performance in 2013 88.2% (87.4-89.1) n=5720 than patients from communities with a median income above $69,000 91.5% (91.0-92.1) n=10,076. In 2008, the mean performance rate for <$36K was 79.7% (78.7-80.8) n=5507 and increased by 8.5% by 2013. Patients that resided in communities with median incomes above $69K experienced a 8.1% gain in performance between 2008 and 2013.

Proportion of residents with no high school degree in zip code
Patients that resided in communities at time of diagnosis with the lowest proportion of no high school degree (<7%) had higher rates of performance in 2013 91.0% (90.4-91.6) n=9771) than patients from communities with the highest proportion of patients with no high school degree (>21%) 89.4% (88.7-90.1) n=7302. Likewise, the performance increase from 2008 was higher for patients from

communities with the highest proportion of no high school degree (>21%) 9.5% gain.

Facility Type
Facility type was assessed by programs CoC-accreditation status. Facility types include Comprehensive Community Cancer Programs, Integrated Network Cancer Programs, Community Cancer Programs and by Academic/Research programs. Since 2008, patients at community hospitals experienced the largest gain in performance (+10.8%), whereas the performance rate in 2008 for patients at academic centers was 86.0% (85.3-86.7), representing a 6% gain. Patients that were treated at teaching/research hospitals experienced higher performance rates 92.0% (91.5-92.5) n=11,214 to those treated at comprehensive community centers 89.3% (88.8-89.7) n=21,302 in 2013. Patients treated at smaller community hospitals had lower performance rates at 86.4% (85.5-87.3) n=5902.

Census Region
There was very little regional variation by Census Region in 2013. Patients that resided in the Midwest Census Region had a performance rate of 90.4% (89.8-90.9) n=10,754 in 2013, Pacific 90.2% (89.3-91.0) n=4748, West 90.1% (88.7-91.5) n=1700, South 88.9% (88.4-89.4) n=15,614, and Northeast 89.9% (89.3-90.5) n=8651. Patients that resided in the Pacific Census Region saw the largest performance gain since 2008 79.2% (78.1-80.4) n=4695 representing a 11.2% increase. By contrast, the smallest increase in performance was in the Northeast region, 83.1% (82.3-84.0) n=8036 in 2008.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
The American College of Pathologists (1999) recommended that a minimum of 12 lymph nodes be examined to accurately identify AJCC Stage III colon cancer. The American Joint Committee on Cancer (5th edition) indicated that it was desirable to obtain at least 12 lymph nodes in radical colon resections (1997). The AJCC (6th edition) modified this recommendation to obtain at least 7-14 lymph nodes, but included rectal resections among the procedures associated with this numeric recommendation. By its 7th edition, citing data from NCI/SEER, clearly noted the positive relationship between the number of nodes pathologically examined and patient survival.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Compton CC, Fielding LP, Burgart LJ, et al. Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000; 124:979-994. 2. Fleming ID, Cooper JS, Donald EH, et al (eds). AJCC Cancer Staging Manual, Fifth edtion. Lippincott-Raven 1997, p. 84. 3. Greene FL, Page DL, Fleming ID, et al (eds.) AJCC Cancer Staging Manual, Sixth edition. Spinger 2002, p. 114. 4. Edge SB, Byrd DR, Compton CC, et al (eds.) AJCC Cancer Staging Manual, Seventh edition. Spinger 2010, p. 153

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Colorectal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Disparities

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.facs.org/~/media/files/quality%20programs/cancer/ncdb/measure%20specs%20colon_03312015.ashx

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Since the last endorsement, this measure has been updated to include any tumor diagnosis, no longer limited to first or only malignant primaries. Due to advances in medical oncology, drug development, surgical techniques and radiotherapeutic techniques, more patients are living longer and being cured of their cancer only to develop a subsequent malignancy. For this reason when patients are being operated on for curative intent, we are holding these patients to the standard of care of a threshold minimum lymph node removal established in the Commission on Cancer Quality Measures. We have also allowed for exclusion of cases based on perforation to the primary site and acute obstructions which may lead to emergency interventions and lack of a lymph node yield.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
>=12 regional lymph nodes pathologically examined.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
This measure is assessed based on data submitted in standard North American Association of Central Cancer Registries (NAACCR), data is reported back to programs reporting annual or quarterly compliance based on the date of diagnosis.

S.6. **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, 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.
Regional Lymph Nodes Examined [NAACCR Item#830] = 12-90

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Include, if all of the following characteristics are identified:
Age >=18 at time of diagnosis
Primary tumors of the colon
Epithelial malignancy only
AJCC Stage I, II, or III
Surgical resection performed at the reporting facility

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Surgical Procedure of the Primary Site at This Facility [NAACCR Item#670] = 30-80

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude, if any of the following characteristics are identified:
Age <18; non-epithelial and non-invasive tumors; metastatic disease (AJCC Stage IV); not treated surgically at the reporting facility; perforation of the primary site; acute obstruction

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20colon.ashx

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No stratification applied

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)
This measure score is calculated by dividing the numerator cases by denominator eligible cases.

Denominator eligible cases are assessed in a step-wise fashion:
First include diagnosis of colon cancer,
male or females,
adult patients,
malignant tumors
epithelial tumors based on AJCC 7th edition staging
Exclude clinical or pathologic evidence of in-situ disease
Exclude clinical or pathologic evidence of metastatic disease
All or part of first course of treatment at reporting facility
surgically treated at this facility

These cases are included in the denominator

Then numerator cases are assessed from denominator eligible cases
The number of regional lymph nodes examined are 12-90
The number of nodes examined is greater or equal to the number of positive lymph nodes.

The measure score is calculated by the numerator divided by the denominator.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
No sampling

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
NA

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
All cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospital cancer registry data, reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0225_MeasureTesting_MSF5.0_Data.doc,0225_MeasureTesting_12RLN_04012016.doc

NATIONAL QUALITY FORUM

NQF #: 0225 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
This measure has been implemented by the ACoS CoC since 2007 across all CoC-accredited cancer programs, and reports on approximately 37,800 cases per year to almost 1,400 cancer programs.

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)*:
Cancer registry case records reported to the NCDB are reviewed annually, annualized hospital performance rates are provided back to CoC accredited cancer programs via the CoC´s Cancer Program Practice Profile Report (CP3R) using the denominator and numerator criteria documented in response to items 2a1.3 and 2a1.7, respectively, in the Specifications section. (http://www.facs.org/cancer/ncdb/cp3r.html)

2a2.3 **Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The CoC as been able to track an upward trend in cancer program compliance with this measure. For cases diagnosed in 2008 the mean program performance rate is 80.4%, while the median was 83.3%. These rates continue to document an increase in aggregate performance rate over time. In 2007, the median performance rate was almost 79%, and mean performance rate was 75%. Analysis of data from 2009 indicate the mean program performance rate has increased to 81.5%, with a median value of 85.7%. Low performance outliers have been observed continuously over time. For example, in 2008 2.5% (n=34) of programs had a performance rate below 41%.

The mean performance rate for this measure in 2013 was: 89.7% (89.4-90.0) IQR=84-97% n=41,546. The minimum hospital-level performance rate is 0% with a 100% maximum in all years assessed 2008-2013. 189 programs in 2013 reported compliance <=75.0%, representing the lowest 10th percentile of programs.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the**

evidence cited in support of the measure focus (*criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
See 2a2.1. This measure has been implemented across all CoC-accredited cancer programs and subject to local review by standing committies of these hospitals and site surveyors at the time of accreditation site visits.

During Commission on Cancer Survey Site visits in 2009 and 2010, surveyors validated not more than 25 charts.
During 2009 – 391 accredited sites were reviewed, including 5,712 charts. This included an average of 14 charts per survey (IQR 6-22). 322 of these charts were colon cases; representing 10.3% of measure eligible cases.
During 2010- 423 accredited sites were reviewed, including 6,752 charts. This was based on an average of 14 charts per survey (IQR 6 – 22). 382 of these charts were colon cases; representing 10.6% of measure eligible cases.

2b2.2 **Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
Performance rates are reviewed and discussed, randomly selected charts are reviewed by the site surveyor to ascertain the completeness and validity of the data recorded in the local cancer registry and reported to the NCDB and included in the CP3R reporting application.

Major areas of review completed by site surveyors included but were not limited to, confirmation of timing of adjuvant therapy, documentation of treatment recommended but not received, assessment of missing and incomplete tumor characteristics.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
This measure has a high degree of user acceptability, the measure denominator and numerator are viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Assessment of timing for chemotherapy for cases in which there were positive lymph nodes found where treatment was significantly later than expected (>90 days after diagnosis) this measure had the highest concordance with 88.9% in 2006 diagnoses and 81.8% for 2007 cases. There was 88.5% and 92.4% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The NCDB collects all diagnosed cases within cancer programs. The measure exclusions as described in the specifications are the inverse of the measure inclusion criteria. Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure. These are established to ensure patients included in the measure assessment meet the evidence based criteria. In 2013, 41,546 colon cases were included in this measure.

The exception to this is the measure exclusions of: Perforation of the primary site and Acute obstruction. These exclusions have been added to account for emergent cases.
2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed. For all cases applicable to this measure, in 2012 -2013, 16 cases were excluded for Perforation of the primary site and 26 cases were excluded based on acute obstructions. This is a total of 42 of 41546 eligible cases excluded from this measure.

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
Measure exclusions were input for n=42 cases representing <0.01% of eligible cases and does not affect estimated performance rates for this measure.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not applicable

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Differences in data performance was described in performance gaps.

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:**

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
This measure was not specified to report stratified performance rates, however the CoC´s recently released (2011) "real clinical time" Rapid Quality Reporting System (RQRS) (http://www.facs.org/cancer/ncdb/rqrs.html) reports back measure-specific performance rates by a number of strata, eg. patient age, sex, ethnicity, insurance status, and area-based SES. RQRS hosts a prospective treatment alert system, and so performance rates are both high and consistent with clinical expectation, however room for potential improvement remains. In a comparative analysis of 16 NCI/NCCCP pilot sites using RQRS with a comparative group of 25 other CoC-accredited

cancer programs also using RQRS revealed that at NCCCP cancer programs female patients more frequently received adjuvant chemotherapy (88.2%) than did males (86.3.9%). Comparative rates from the 25 non-NCCCP programs showed almost no difference in performance rates for this measure based upon patient sex. However, there was an almost 5% difference between the proportion of patients under the age of 50 having 12+ lymph nodes examined, compared to patients 70 or older (89.1% v 85.5%). Analysis from cases diagnosed 2008-2010.
Additional disparities data was presented in section 1.b. of this application.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

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)
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)*
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.
The ACoS/CoC implementation of this measure is framed around the feasibility of data collection and reporting considerations. Cancer registries in the United States depend on a multitude of information sources in order to completely abstract case records and be in compliance with State, Federal and private sector accreditation requirements. Commission on Cancer Standards require case abstracting to be performed by a Certified Tumor Registrars (CTRs). CTRs must pass an exam and maintain continuing education. In the past decade, great strides have been made within the cancer registration community in terms of electronic capture of registry data from electronic pathology systems and electronic health records. However, until EHR systems are universally implemented in the US and fully integrated within hospital-level cancer registry systems, registry data will depend upon some level of human review and intervention to ensure data are complete and accurately recorded. Robust data quality edits are applied to the data at all levels of cancer data abstraction and processing. These edits standardize coded information and ensure its accuracy.

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.
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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
1) The infrastructure to monitor compliance with this measure has been in place since 2005 to assess and feed-back to approximately 1,500 Commission on Cancer (CoC) accredited centers performance rates for this measure. CoC accredited cancer programs account for 70-80% of patients affected by this measure. This measure is currently reported to CoC accredited programs through the National Cancer Data Base (NCDB) using the Cancer Program Practice Profile Report (CP3R) web-based audit and feed-back reporting tool. The CP3R is generally described at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. In addition, this measure is also reported to over 1000 cancer programs participating in its "real clinical time" feedback reporting tool through its Rapid Quality Response System (RQRS). An overview of the RQRS is available at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Both of these reporting tools have been utilized in the cancer registry community and will not produce an undue burden on the data collection network.

2) The data for this measure are key elements already collected in all hospital registries. This measure has been reviewed using cancer registry data. The CoC data demonstrates variation in the measure. The measure is readily implemented.

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)*.

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.

Planned	Current Use (for current use provide URL)
Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	Public Reporting Pennsylvania Health Care Quality Alliance http://www.phcqa.org/ Regulatory and Accreditation Programs Commission on Cancer https://www.facs.org/quality-programs/cancer/coc/standards Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Commission on Cancer, National Cancer Data Base https://www.facs.org/quality%20programs/cancer/ncdb Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/manual-qopi-measures

4a.1. For each CURRENT use, checked above, provide:

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

4.1.a Public Reporting

Pennsylvania Health Care Quality Alliance (PHCQA)
About: PHCQA is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. This organization works to develop a consensus-driven, statewide approach to hospital quality measurement that is supported by quality of care data from a variety of public data sources. Commission on Cancer (CoC) accredited cancer programs in Pennsylvania may elect to voluntarily report their estimated performance rates through this program. Currently 60 of 73 Pennsylvania programs are participating.

4.1.d Regulatory and Accreditation Programs
Commission on Cancer Accreditation
The CoC standards have a requirement for programs to meet expected estimated performance rates for accountability measures approved through the CoC (Standard 4.5), programs are required to meet 85% compliance with this measure, either through the estimated performance rate point estimate or the 95% confidence interval. If this benchmark was not met then programs were required to develop action plans to address this issue. There are over 1500 CoC-accredited cancer programs representing approximately 70% of newly diagnosed cancer cases annually.

4.1.f Quality Improvement with Benchmarking

Commission on Cancer, National Cancer Data Base
Purpose: The National Cancer Data Base (NCDB) provides a benefit for CoC-accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP).

CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer and is described https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. This application is available to over 1500 CoC-accredited cancer programs

CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cqip

RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to quality of cancer care measures for breast and colorectal cancers -https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

Quality Oncology Practice Initiative
In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

4a.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?*)

4a.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.)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

2008: 81.8 (81.4 – 82.2) N = 39,910
2009: 84.6 (84.3 – 85.0) N = 38,578
2010: 86.3 (85.9 – 86.6) N = 36,145
2011: 87.6 (87.3 – 88.0) N= 35,738
2012: 88.1 (87.8 – 88.4) N= 42,570
2013: 89.7 (89.4 – 90.0) N= 41,546

Patient demographics, including ethnicity and age were considered. The trends reveal an increase in performance across all ethnicity and age groups, although the rates within vary. Non-hispanic whites, Non-hispanic blacks, Asian Pacific Islanders, Hispanics, and Non-hispanic Blacks had similar performance rates in 2013; Non-hispanic whites at 89.9% (95% CI: 89.6-90.3) n=31,698 in 2013, non-Hispanic Blacks at 88.6% (95% CI: 87.7-89.6) n=4752, Asian Pacific Islanders 90.6% (88.9-92.4) n=1064, and Hispanics 89.7% (95% CI: 88.4-90.9) n=2177. Between 2008 and 2012, performance rates increased in all ethnic groups. Patients under the age of 50 at diagnosis had higher performance rates in 2013 at 94.5% (93.7-95.2) n=3164, compared to patients aged 50 to 59 90.0% (89.3-90.7) n=6776, patients aged 60-69 89.7% (89.1-90.3) n=9762, patients aged 70-79 89.2% (88.7-89.8) n=11,196, and patients aged >79 88.6% (88.0-89.2) n=10,648. Since 2008, patient above age 49 saw a relatively equal gain in performance with the measure ~9%. Patients age 18-49 saw only a 5.2% increase given the already high performance 89.3% (88.3-90.4) n=3360 in 2008.

Geographic variation
Geographic regional variation was limited by 2013. Patients that resided in the Midwest Census Region had a performance rate of 90.4% (89.8-90.9) n=10,754 in 2013, Pacific 90.2% (89.3-91.0) n=4748, West 90.1% (88.7-91.5) n=1700, South 88.9% (88.4-89.4) n=15,614, and Northeast 89.9% (89.3-90.5) n=8651. Patients that resided in the Pacific Census Region saw the largest performance gain since 2008 79.2% (78.1-80.4) n=4695 representing a 11.2% increase. By contrast, the smallest increase in performance was in the Northeast region, 83.1% (82.3-84.0) n=8036 in 2008.

4b.2. **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.**
NA

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
This measure, as specified, is unlikely to be systematically susceptible to under-reporting due to the integral dependence of the measure upon information routinely documented and reported following pathologic examination of colon tissue specimens.

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

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 completely harmonized?

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):** Commission on Cancer, American College of Surgeons
Co.2 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-
Co.3 **Measure Developer if different from Measure Steward:** Commission on Cancer, American College of Surgeons
Co.4 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-

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 Developers: Christopher (Chris) Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); Richard Swanson, MD, FACS (Partners Health Care, Boston MA); Peter Enzinger, MD (Dana Farber Cancer Institute, Boston MA); Elin Sigurdson, MD, FACS (Fox Chase Cancer Center, Philadelphia PA); Mitchell Posner, MD, FACS (University of Chicago, Chicago IL); Anthony Robbins, MD, PhD (American Cancer Society)

The current Measure workgroup includes:
Charles Cheng MD, FACS (Fox Valley Surgical Associates, Appleton, WI), Daniel McKellar, MD, FACS (Wayne Healthcare, Greenville, OH), David Jason Bentrem, MD (Northwestern Memorial Hospital, Chicago, IL), Karl Bilimoria, MD, FACS (Northwestern Univ/Feinberg Sch of Med, Chicago, IL), Lawrence Shulman MD (University of Pennsylvania, Philadelphia, PA), Matthew A Facktor, MD FACS (Geisinger Medical Center, Danville, PA), Ted James (University of Vermont, Burlington, VT)

This panel meets at least once a calendar year to review quality measures currently supported and implemented by the ACoS Commission on Caner and to investigate and consider/review development of possible new measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 06, 2007
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?** 11, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 0377
Measure Title: Hematology: Myelodysplastic Syndrome (MDS) and Acute Leukemias: Baseline Cytogenetic Testing Performed on Bone Marrow
Measure Steward: American Society of Hematology
Brief Description of Measure: Percentage of patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) or an acute leukemia who had baseline cytogenetic testing performed on bone marrow
Developer Rationale: Cytogenetic testing is an integral component in calculating the International Prognostic Scoring System (IPSS) score. Cytogenetic testing should be performed on the bone marrow of patients with MDS in order to guide treatment options, determine prognosis, and predict the likelihood of disease evolution to leukemia.

For acute leukemias:
In addition to establishing the type of acute leukemia, cytogenetic testing is essential to detect chromosomal abnormalities that have diagnostic, prognostic, and therapeutic significance. Performing cytogenetic analysis on patients with AML identifies a subgroup of patients where further molecular genetics testing is indicated.

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.
National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute Myeloid Leukemia. Version 1, 2016.

Numerator Statement: Patients who had baseline cytogenetic testing performed on bone marrow
Denominator Statement: All patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) or an acute leukemia
Denominator Exclusions: For Registry:
Documentation of medical reason(s) for not performing baseline cytogenetic testing (eg, no liquid bone marrow or fibrotic marrow)
Documentation of patient reason(s) for not performing baseline cytogenetic testing (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above)
Documentation of system reason(s) for not performing baseline cytogenetic testing (eg, patient previously treated by another physician at the time cytogenetic testing performed)

Measure Type: Process
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Aug 09, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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:

- The evidence for this measure is based on a clinical practice guideline from the National Comprehensive Cancer Network (NCCN) for cytogenetic testing patients with MDS and AML. **Level of evidence: Category 2A.**
  - NCCN describes category 2A as based on "lower level evidence, there is uniform NCCN consensus that the intervention is appropriate."
- The developer states that the quality of the body of evidence supporting the guideline recommendation was categorized as "lower- level evidence" which may include non-randomized trials; case series; or when other data are lacking, the clinical experience of expert physicians.
- In 2012, the Committee was concerned that the literature cited and rationale provided by measure authors focused mainly on the use of cytogenetics in MDS and its evolution to acute myelogenous leukemia (AML) and did not include much information on *de novo* AML. Although much of the literature presented in the application was based on retrospective reviews, there was some prospective randomized literature in AML that was stratified based on prognostic factors (including cytogenetics) to indicate that cytogenetic abnormalities predict outcome. However, this measure is based mainly on a consensus guideline from the National Comprehensive Cancer Network (NCCN). The authors graded the literature as 2A based on lower level evidence.

Updates: The developer provided updates to the guidelines – no changes were made to the evidence.

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:**

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as lower-level evidence (Box 6) → Low (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review.*

- *Is the SC aware of higher-level evidence to support this measure?*
- *Does the SC think there is a need to repeat the discussion and voting on 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.

The developer provided the following performance data from 1/1/2014 to 12/31/2014:

- **Registry Performance Rates:**
  - Mean: 95.09%
  - Minimum: 22.22%
  - Maximum: 100.00%
- **PQRS Average Performance Rates:**
  - 2010- 88.8%
  - 2011- 94.6%
  - 2012- 95.6%
  - 2013- 87.0% - Beginning in 2015, PQRS began imposing payment penalties for non-participants based on 2013 performance. For 2013, 5.7% of eligible professionals participating reported on MDS: Baseline Cytogenetic Testing. As a result, performance rates may not be nationally representative.
- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).
- The developer provided additional data from the literature.

Disparities:

- The developer stated that federal reporting programs have not yet made disparities data available to analyze and report. Disparities data from the measure as specified is required for endorsement maintenance; this measure has been endorsed since 2008.
- The developer states they are not aware of any literature that addresses disparities in patients with ACL and MDS receiving baseline cytogenetic testing.

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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 to Support Measure Focus

Comments:

**Precision medicine in form of cytogenetic testing for new diagnosis AML or MDS remains critical for diagnosis, prognosis and therapeutic options. This maintenance process measure originally endorsed 2008, last endorsed 2012, has unchanged evidence, remains consensus based from NCCN graded 2A (lower level evidence, there is uniform NCCN consensus that the intervention is appropriate). No need to repeat the discussion for voting on evidence. After literature search, SC member not aware of higher level evidence. Prospective RCT's would be limited in this setting whereas studies on use of longitudinal marrows or cytogenetic risk stratification exists. Newer molecular, FISH diagnostic panels for MDS, acute leukemia complement but do not remove necessity of cytogenetics testing at time of diagnosis. Given rapidly evolving field, real time risk stratification with molecular testing needed but may not be realistic request for this measure. Would be ideal if measure developer could specify additional molecular testing to cytogenetics or at least add

FISH to current measure (NCCN grade 2A) or molecular testing for intermediate risk, normal karyotype, good risk with c-kit, as well as develop composite measure that takes into account over performance of follow up bone marrows after diagnosis that may occur. This may make measure more relevant in community practices.**
Yes to all.

1b. Performance Gap
Comments:
Data supports quality problem. Though low sample size, range of physician performance from 0.26 to 1.00 suggests meaningful variation across physicians' performance, thus, opportunity for improvement from data available 2014. Disparities data not available.
Yes; no data on subgroups and there are disparities.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Registry

Specifications:

- This is a clinician-level measure
- The numerator includes patients who had baseline cytogenetic testing performed on bone marrow
- The denominator includes all patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) or an acute leukemia
- Denominator exclusions include:
    - Documentation of medical reason(s) for not performing baseline cytogenetic testing (e.g., no liquid bone marrow or fibrotic marrow)
    - Documentation of patient reason(s) for not performing baseline cytogenetic testing (e.g., at time of diagnosis receiving palliative care or not receiving treatment as defined above)
    - Documentation of system reason(s) for not performing baseline cytogenetic testing (e.g., patient previously treated by another physician at the time cytogenetic testing performed)
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The calculation algorithm is provided.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability was conducted on a sample size from 2008 and 2009 that included 29 acute leukemia patient records and 31 MDS patient records from two hematology practice sites. Chart and data auditing occurred in 2010. The developer provided the percent agreement and kappa statistic (95% CI) for 60 charts: :
    - Overall Reliability: 98.3%, 0.9138 (0.7469 – 1.0000)
    - Denominator Reliability: 100.0%

- Numerator Reliability: 98.3%, 0.9138, (0.7469 – 1.0000)
- Exceptions Reliability: 100.0%

Describe any updates to testing:

- Reliability of the measure score was not presented in prior submission(s), reliability testing of the measure score has been conducted this review.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 2014 Registry data from PQRS. A total of 151 physicians reported on this measure in 2014. Of those, 67 physicians had 1,432 patient charts with all the required data elements and a **minimum of 10 quality reporting events**. The **average number of quality reporting events** (after exceptions were removed) was **21.0**.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability.

Results of reliability testing:

- Reliability at the at the minimum level of quality reporting events (10) was **0.68** and **0.82** at the average number of quality events (21.0).

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Appropriate method used (Box 5) →High/moderate reliability statistic and scope (Box 6) →Moderate

Questions for the Committee:

- *Is the test sample of 67 physicians adequate to generalize for widespread implementation?*
- *Is it likely this measure can be consistently implemented?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing:

- Additional empirical validity testing of the measure score has been conducted since the last review of this measure.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

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

Validity testing method:

- Face validity was assessed using a panel of experts with representation from the ASH Committee on Quality.

Validity testing results:

- **94%** of the respondents either agreed or strongly agreed that this measure can accurately distinguish good and poor quality.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer noted several exclusions, as follows:
  - Documentation of medical reason(s) for not performing baseline cytogenetic testing (e.g., no liquid bone marrow or fibrotic marrow)
  - Documentation of patient reason(s) for not performing baseline cytogenetic testing (e.g., at time of diagnosis receiving palliative care or not receiving treatment as defined above)
  - Documentation of system reason(s) for not performing baseline cytogenetic testing (e.g., patient previously treated by another physician at the time cytogenetic testing performed)
- The developer reported that there were a total of 17 exceptions reported amongst the 67 physicians with the minimum (10) number of quality reporting events. The average number of exceptions per physician was 0.25 and overall exception rate was 1.2%.
- Without the exclusions "the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives."
- The developer also stated that they recommend physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. ASH also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Questions for the Committee:

- *Are the results from the exclusion analysis a threat to validity?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer calculated measures of central tendency, variability, and dispersion. Based on a sample of 67 physicians:
    - Mean performance rate is 0.93
    - Median performance rate is 100%, and the mode is 1.00
    - Standard deviation is 0.13
    - Range of the performance rate is 0.74 , with a minimum rate of 0.26 and a maximum rate of 1.00
    - Interquartile range is 0.08 (0.92 – 1.00)

Question for the Committee:

- *Does a sample of 67 physicians adequately identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- The developer stated test was not performed for this measure.

2b7. Missing Data

- The developer stated data are not available to complete this testing.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Threats to validity mostly assessed (Box 2) →Empirical validity testing (Box 3)→ Face validity assessed (Box 4)→ Agreement measure can be used to distinguish quality (Box 5)→Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Data elements clearly defined. Calculation algorithm clear. Measure that can be implemented consistently --Limited sample size for rare diseases.

Yes.

Specifications consistent with evidence.

They are consistent.

2a2. Reliability Testing

Comments:

Inter rater reliability overall 98.3% (previous). Test sample 67 physicians. Measure score with signal to noise analysis reliability score .82 (average # of quality events)- MODERATE.

Yes

2b2. Validity Testing

Comments:

Face validity minimal but acceptable with substantial agreement 94% and no threats to validity, potential of bias – MODERATE

Yes validity tested.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Low overall exception rate 1.2%. Exclusion analysis not a threat to validity.

No risk adjusted method. the sample is adequate. moderate validity

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data collection burden due to manual chart abstraction requirement.
- The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Data elements routinely generated, used for care delivery and available by EHR - Moderate to High.
Should be a data field in the EHR.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details :

- Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS). The developer stated that CMS announced that there are plans to make all PQRS individual EP level PQRS measures available for public reporting on Physician Compare in late 2017.
- ASH MDS PIM - an MOC Practice Assessment activity
- This measure has been endorsed since 2008 - per NQF criteria, performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results :

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation:

- The developer reports no additional difficulties or unexpected findings or benefits, apart from those included throughout the submission form.

Potential harms: The developer reports no unintended consequence were noted.

Feedback:

- In 2012, the Committee made the following recommendations:
    - This measure is becoming outdated, as diagnostic panels for MDS and acute leukemias rely heavily upon molecular panels and FISH in addition to standard cytogenetics. The responsibility for these assays is also divided between pathologists (who have no ongoing relationship with patients) and hematologists, who provide ongoing care. The Steering Committee recommended that the measure developer consider specifying this measure in the future to capture FISH and other tests.
    - The Steering Committee recommended the measure developer consider specifying the measure to capture patients with MDS, acute myelogenous leukemia and acute lymphoblastic leukemia. The Committee believed that karyotypic data, stratified appropriately, might provide a way to make major therapeutic decisions with respect to the patient population.
    - The developer responded that they would look to address these concerns in future iterations of the measure.

***Questions for the Committee**:*

- *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?*
- *Did the developer address the concerns of the 2012 Committee?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Late 2017 public reporting planned.

Not currently reported to the public; accountability is in the context of clinical trials.

Criterion 5: Related and Competing Measures

Related or competing measures

No related or competing measures were identified.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

NQF #: 0377 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)***:**

The measure focus is patients with a diagnosis of MDS or Acute Leukemias and performance of baseline cytogenetic testing on bone marrow.

For MDS, cytogenetic testing is an important component to calculating the International Prognostic Scoring System (IPSS) score. Cytogenetic testing should be performed on the bone marrow of patients with MDS in order to guide treatment options, determine prognosis, and determine any possible transition to leukemia.

For acute leukemias, cytogenetic testing is critical to both identify the type of acute leukemia and detect chromosomal abnormalities which contain diagnostic, prognostic, and therapeutic information.

Further molecular genetic testing is indicated when specific cytogenetic abnormalities are identified.

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2012.

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

The guidelines developed by NCCN provide evidence for cytogenetic testing both patients with MDS and AML. Our measure focuses on patients with MDS and AML.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** The description of the evidence review in the guideline did not address the overall quantity of studies in the body of evidence however the AML guidelines reference 79 articles and the MDS guidelines reference 160 articles.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of the body of evidence supporting the guideline recommendation is summarized according to the NCCN categories of evidence and consensus as being based on "lower-level evidence". Lower-level evidence is later described as evidence that may include non-randomized trials; case series; or when other data are lacking, the clinical experience of expert physicians.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Although there is no explicit statement regarding the overall consistency of results across studies in the guidelines supporting the measure, the recommendation received uniform NCCN consensus that the intervention is appropriate.

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit -*

benefit over harms):

For MDS:

Cytogenetic testing is an integral component in calculating the International Prognostic Scoring System (IPSS) score. Cytogenetic testing should be performed on the bone marrow of patients with MDS in order to guide treatment options, determine prognosis, and predict the likelihood of disease evolution to leukemia.

For acute leukemias:

In addition to establishing the type of acute leukemia, In addition to defining some types of acute leukemia, cytogenetic analysis detects chromosomal abnormalities which contain diagnostic, prognostic and therapeutic significance.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: A panel of experts with members from each of the NCCN Member Institutions develops the NCCN Guidelines. Specialties that must be included on a particular panel are identified before that panel is convened but also evolve as the standard of care changes over time. This multidisciplinary representation varies from panel to panel. The NCCN Guidelines Panel Chairs are charged with ensuring that representatives of all treatment strategies are included. Many of the panels also include a patient representative, especially when issues of long-term care and patient preference are paramount in the panel´s considerations.

NCCN publishes individual disclosures of potential conflicts of interest for panel members, NCCN Guidelines staff, and NCCN senior management. Relationships disclosed include research funding, participation in advisory groups, participation in speakers' bureaus, employment, and equity or patent ownership. Beginning in 2010, the NCCN Board of Directors has directed that panel members compensation from external sources be less than published thresholds. These thresholds are <= $20,000 from a single entity and <= $50,000 in aggregate from any source.

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Panel members identify the level of evidence supporting each recommendation. These categories are:

•Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

•Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

1c.13 Grade Assigned to the Body of Evidence: Category 2A

1c.14 Summary of Controversy/Contradictory Evidence: No controversy or contradictory evidence with regard to the importance of identifying normal tissue dose constraints.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

For MDS:

Bone marrow aspiration with Prussian blue stain for iron and biopsy are needed to evaluate the degree of hematopoietic cell maturation abnormalities and relative proportions, percentage of marrow blasts, marrow cellularity, presence or absence of ringed sideroblasts (and presence of iron per se), and fibrosis. Cytogenetics for bone marrow samples (by standard karyotyping methods) should be obtained because they are of major importance for prognosis (Category 2A). (~~NCCN MDS 2012~~) (NCCN MDS 2016)

Significant independent variables for determining outcome for both survival and AML evolution were found to be marrow blast percentage, number of cytopenias, and cytogenetic subgroup (good, intermediate, poor). The percentage of marrow blasts was divisible into four categories: 1) less than 5%, 2) 5% to 10%, 3) 11% to 20%, and 4) 21% to 30% (Category 2A). (~~NCCN MDS 2012~~) (NCCN MDS 2016)

~~A chromosome abnormality confirms the presence of a clonal disorder aiding the distinction between MDS and reactive causes of dysplasia, and in addition has major prognostic value. Cytogenetic analysis should therefore be performed for all patients in whom a bone marrow examination is indicated (BCSH, 2003).~~

Cytogenetic analysis should be performed on all patients with suspected MDS to confirm the diagnosis, inform management options and provide prognostic information. Cytogenetic analysis should be performed on at least 25 metaphases and should be reported in accordance with the International System for Human Cytogenetic Nomenclature Recommendations (Schaffer *et al* 2009). Identification of clonal chromosomal abnormalities has become essential for the application of international prognostic scoring systems (such as the International Prognostic Scoring System [IPSS] and the revised IPSS [IPSS-R]). A new comprehensive cytogenetic scoring system has been incorporated into the IPSS-R (Schanz, *et al* 2012). In addition, identification of a specific cytogenetic abnormality may provide a marker for assessing response to therapy. In patients where conventional marrow cytogenetic analysis is not possible ('dry tap') or has failed, fluorescence *in situ* hybridization analysis of bone marrow or peripheral blood films for selected cytogenetic anomalies (for instance monosomy 7, deletion of 5q, trisomy 8) may help provide diagnostic and prognostic evaluation (Evidence levels 2B,C). (BCSH, 2014)

Acute Lymphoblastic Leukemia:

Hematopathology evaluations should include morphologic examination of malignant lymphocytes using Wright-Giemsa-stained slides and hemtoxylin and eosin (H&E)-stained core biopsy and clot sections, comprehensive immunophenotyping with flow cytometry, and assessment of cytogenetic or molecular abnormalities. Identification of specific recurrent genetic abnormalities is critical for disease evaluation, optimal risk stratification, and treatment planning. (Category 2A Recommendation) (NCCN, 2015)

For AML:

Although cytogenetic information is usually unknown when treatment is initiated in patients with de novo AML, karyotype represents the single most important prognostic factor for predicting remission rate, relapse, and overall survival. Therefore, the importance of obtaining sufficient samples of marrow or peripheral blood blasts at diagnosis for this analysis cannot be overemphasized (Category 2A Recommendation).

The importance of obtaining adequate samples on marrow or peripheral blood at diagnosis to do full karyotyping as well as FISH probes for the most common abnormalities cannot be overemphasized. In addition to basic cytogentic analysis, new molecular markers are helping to refine prognostics groups particularly in patients with a normal karyotype (Category 2A Recommendation) (NCCN AML 2012) (NCCN AML 2016).

1c.17 Clinical Practice Guideline Citation: ~~National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2012.~~

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.

~~British Committee for Standards in Haematology (BCSH). Guidelines for the diagnosis and therapy of adult myelodysplastic syndromes. British Journal of Haematology. 2003; 120: 187-200~~

Killick SB, Carter C, Culligan D, Dalley C, Das-Gupta E, Drummond M, Enright H, Jones GL, Kell J, Mills J, Mufti G, Parker J, Raj K, Sternberg A, Vyas P, Bowen D, and British Committee for Standards in Haematology. Guidelines for the diagnosis and management of adult myelodysplastic syndromes. British Journal of Haematology, 2014, 164, 503–525

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute lymphoblastic leukemia. Version 2, 2015. Available at: http://www.nccn.org/professionals/physician_gls/f_guidelines.asp.

~~National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute myeloid leukemia. Version 2, 2012.~~

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute myeloid leukemia. Version 1, 2016.

1c.18 National Guideline Clearinghouse or other URL: www.nccn.org

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: same as in 1c.10

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: same as 1c.12

1c.23 Grade Assigned to the Recommendation: Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: Moderate **1c.26** Quality: Moderate**1c.27** Consistency**:** Moderate

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0377_Evidence_form_FINAL-635933099681074204.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
For MDS:
Cytogenetic testing is an integral component in calculating the International Prognostic Scoring System (IPSS) score. Cytogenetic testing should be performed on the bone marrow of patients with MDS in order to guide treatment options, determine prognosis, and predict the likelihood of disease evolution to leukemia.

For acute leukemias:
In addition to establishing the type of acute leukemia, cytogenetic testing is essential to detect chromosomal abnormalities that have diagnostic, prognostic, and therapeutic significance. Performing cytogenetic analysis on patients with AML identifies a subgroup of patients where further molecular genetics testing is indicated.

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.
National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute Myeloid Leukemia. Version 1, 2016.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014

Registry Performance Rate:
Mean: 95.09%
Minimum: 22.22%
Maximum: 100.00%

PQRS Experience Report

2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on Hematology: Myelodysplastic Syndrome (MDS) and Acute Leukemias: Baseline Cytogenetic Testing Performed on Bone Marrow were:
Average Performance Rate:
2010- 88.8%
2011- 94.6%
2012- 95.6%
2013- 87.0%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 5.7% of eligible professionals participating reported on MDS: Baseline Cytogenetic Testing. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends.
Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

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.**
Thirty percent of general pathologists and 22% of hematopathologists would not include bone marrow aspirate and cytogenetics in initial testing of a neutropenic patient. Most practitioners tested appropriately for disease classification and prognosis; discrepancies were identified in testing to differentiate MDS from acute myeloid leukemia and testing in post treatment specimens. These results have implications in the management of MDS.

Glauser TA1, Sagatys EM, Williamson JC, Burton BS, Berger C, Merwin P, Sugrue M, Bennett JM. Current pathology practices in and barriers to MDS diagnosis. Leuk Res. 2013 Dec;37(12):1656-61. doi: 10.1016/j.leukres.2013.10.007. Epub 2013 Oct 22.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
While this measure is included in federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
We are not aware of any publications/evidence outlining disparities in patients with Acute Leukemias and MDS receiving baseline cytogenetic testing.

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
In the general population, the incidence rate of MDS is approximately 4.8 per 100,000 people per year. However, MDS occurs more greatly among older people occurring in 29.6 per 100,000 people 70-79 years old, and 55.8 per 100,000 people older than 80. (1)

The 5-year survival rate is 29%. Bone marrow blast percentage, number of cytopenias, and cytogenetics represent major factors determining outcomes for patients with MDS in prognostic models such as the International Prognostic Scoring System (IPSS). In addition to karyotype and MDS subtype, transfusion dependence is a key factor in the WHO-based Prognostic Scoring System (WPSS). The MDS are a group of understudied hematologic disorders, and MDS may be the underlying condition affecting some elderly patients with unexplained anemia. With the current demographic trend, increasing disease morbidity (both incidence and prevalence) is expected in the near future. (2)

Approximately 18,860 people will be diagnosed with acute myeloid leukemia (AML) in 2014, and 10,460 patients will die from the disease. As the population ages, the incidence of AML, along with myelodysplasia, appears to be rising. (3)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1) National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.

2) Ma X. Epidemiology of Myelodysplastic Syndromes. Am J Med. 2012 Jul; 125(7 Suppl): S2–S5. doi: 10.1016/j.amjmed.2012.04.014

3) National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Acute Myeloid Leukemia. Version 1, 2016.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not applicable. Not a PRO-PM.

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 subcriteria 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):*
Cancer, Cancer : Hematologic

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

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.)*
The measure specifications are included as an attachment with this submission. Additional measure details may be found at http://www.hematology.org/Clinicians/Guidelines-Quality/PQRS/503.aspx

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:** NQF0377__I9toI10_conversion.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Supporting guidelines and coding included in the measure are reviewed on an annual basis. However, this annual review has not resulted in any changes for this measure.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had baseline cytogenetic testing performed on bone marrow

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
At least once during the 12 consecutive month measurement period

S.6. **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, 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.
Numerator Definition:
Baseline Cytogenetic Testing: Testing that is performed at time of diagnosis or prior to initiating treatment (transfusion, growth factors, or antineoplastic therapy) for that diagnosis

For Registry:
Report the CPT Category II code: 3155F – Cytogenetic testing performed on bone marrow at time of diagnosis or prior to initiating treatment

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) or an acute leukemia

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Denominator Note:
This measure is to be reported a minimum of once per reporting period for all myelodysplastic syndrome (MDS) and Acute Leukemia patients seen during the reporting period, regardless of when MDS or Acute Leukemia diagnosis was made; the quality action being measured is that baseline cytogenetic testing on bone marrow was performed for each patient with MDS and Acute Leukemia at the time of diagnosis or prior to initiating treatment.

For Registry:
Patients aged >= 18 years
AND
Diagnosis for MDS or acute leukemia - not in remission (ICD-9-CM) [reportable through 09/30/2015]: 204.00, 204.02, 205.00, 205.02, 206.00, 206.02, 207.00, 207.02, 207.20, 207.22, 208.00, 208.02, 238.72, 238.73, 238.74, 238.75
Diagnosis for MDS or acute leukemia - not in remission (ICD-10-CM) [reportable beginning 10/1/2015]: C91.00, C91.02, C92.00, C92.02, C92.40, C92.42, C92.50, C92.52, C92.60, C92.62, C92.A0, C92.A2, C93.00, C93.02, C94.00, C94.02, C94.20, C94.22, C95.00, C95.02, D46.0, D46.1, D46.20, D46.21, D46.22, D46.4, D46.9, D46.A, D46.B, D46.C, D46.Z
AND
CPT codes: 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, 99241, 99242, 99243, 99244, 99245

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
For Registry:
Documentation of medical reason(s) for not performing baseline cytogenetic testing (eg, no liquid bone marrow or fibrotic marrow)
Documentation of patient reason(s) for not performing baseline cytogenetic testing (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above)
Documentation of system reason(s) for not performing baseline cytogenetic testing (eg, patient previously treated by another physician at the time cytogenetic testing performed)

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For the measure Myelodysplastic Syndrome (MDS) and Acute Leukemias – Baseline Cytogenetic Testing Performed on Bone Marrow, exceptions may include medical reasons (eg, no liquid bone marrow or fibrotic marrow), patient reasons (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above), or system reasons (eg, patient previously treated by another physician at the time cytogenetic testing performed). Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Additional details by data source are as follows:

For Registry:
Documentation of medical reason(s) for not performing baseline cytogenetic testing on bone marrow (eg, no liquid bone marrow or fibrotic marrow) - Append modifier to CPT Category II code: 3155F-1P

Documentation of patient reason(s) for not performing baseline cytogenetic testing on bone marrow (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above) - Append modifier to CPT Category II code: 3155F-2P

Documentation of system reason(s) for not performing baseline cytogenetic testing on bone marrow (eg, patient previously treated by another physician at the time cytogenetic testing performed) - Append modifier to CPT Category II code: 3155F-3P

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment or risk stratification.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at*

measure-specific URL identified in S.1.)
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).
2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.
3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: include medical reasons (eg, no liquid bone marrow or fibrotic marrow), patient reasons (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above), or system reasons (eg, patient previously treated by another physician at the time cytogenetic testing performed). If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable. The measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable. The measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not Applicable

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Clinician : Team

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **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. The measure is not a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0377_Evidence_form_FINAL_3-25-16.doc

PREVIOUS MEASURE TESTING

NATIONAL QUALITY FORUM

NQF #: 0377 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
PCPI Testing Project
- Two hematology practice sites representing various types, locations and sizes were identified to participate in testing the measures
- Site A was a hematology group practice with eight physicians that cared for hematology patients. Site B was a large multi-specialty group clinic with 13 physicians that cared for hematology patients.
- Site A had a document retrieval system rather than a full-fledged EHR where data was scanned in and required searching. Site B had a fully functional EHR.
- Both sites were located in urban/suburban regions
- Hematology patient visit volume was 150 per day at site A and 120-150 per day at site B.
- Both sites were instructed to select 120 patient records (20 with acute leukemias and 35 for each of the following diagnoses: MDS, multiple myeloma and CLL).
- At site A the number of patients in practice in 2009 by specialty area was as follows:
  - Myelodysplastic Syndrome (MDS): 145 patients

o Acute Leukemias: 52 patients
• At site B the number of patients in practice in 2009 by specialty area was as follows:
o Myelodysplastic Syndrome (MDS): 15 patients
o Acute Leukemias: 29 patients
• For this measure, the sample size included 60 abstracted patient charts. Site B included more MDS patients in their sample because of difficulties obtaining the required number of acute leukemia patients.
• The measurement period (data collected from patients seen) was between 1/1/2009 through 12/31/2009. Due to an inability to obtain the required number of patient records for acute leukemia and MDS during the specified measurement period, site B also included patients from 2008.
• Chart auditing was performed between 5/17/2010 and 7/15/2010
• Data auditing was performed between 8/2/2010 and 9/14/2010

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)***:**
Data abstracted from patient records were used to calculate inter-rater reliability for the measure.
29 acute leukemia and 31 MDS patient records were reviewed.

Data analysis included:
• Percent agreement
• Kappa statistic to adjust for chance agreement

2a2.3 **Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
N, % Agreement, Kappa (95% Confidence Interval)
Overall Reliability: 60, 98.3%, 0.9138 (0.7469 – 1.0000)
Denominator Reliability: 60, 100.0%, Kappa is noncalculable*
Numerator Reliability: 60, 98.3%, 0.9138, (0.7469 – 1.0000)
Exceptions Reliability: 60, 100.0%, Kappa is noncalculable*

This measure demonstrates almost perfect reliability, as shown in results from the above analysis.

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
The evidence includes both patients with MDS and AML. The NCCN Myelodysplastic Syndrome and Acute Myeloid Leukemia guidelines were developed to provide direction in the evaluation and treatment of these disorders. The population of patients assessed by this measure meet the diagnostic criteria stated in the guidelines.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
An expert panel was used to assess face validity of the measure. This panel consisted of the following 10 members (with specialties listed):

Steven L. Allen, MD (Co-Chair) (hematology/oncology)
William E. Golden, MD (Co-Chair) (internal medicine (IM))
Kenneth Adler, MD (hematology/IM)
Daniel Halevy, MD (nephrology)
Stuart Henochowicz, MD, MBA (IM)
Timothy Miley, MD (hematopathology)
David Morris, MD (radiation oncology)
John M. Rainey, MD (medical oncology)
Samuel M. Silver, MD, PhD (hematology/oncology)
Lawrence Solberg, Jr., MD, PhD (hematology/IM)

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert Work Group and the measures adjusted as needed. Other external review groups (i.e. focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity has been quantitatively assessed for this measure. Specifically, the work group members were asked to empirically assess face validity of the measure. This work group/expert panel consists of 10 members, whose specialties include oncology, hematology, internal medicine, and clinical pathology.

Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

The survey scale is 1-5, where 1=Disagree; 3=Neither Disagree nor Agree; 5=Agree

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
The results of the expert panel rating of the validity statement were as follows: N = 8; Mean rating = 4.75

Percentage in the top two categories (4 and 5): 100%

Frequency Distribution of Ratings
1 - 0
2 - 0
3 - 0
4 - 2
5 – 6

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
PCPI Testing Project
- 60 patient records (29 Acute Leukemia and 31 Myelodysplastic Syndrome) were reviewed for this measure.
- The measurement period (data collected from patients seen) was between 1/1/2009 through 12/31/2009.
- Chart auditing was performed between 5/17/2010 and 7/15/2010.
- Data auditing was performed between 8/2/2010 and 9/14/2010.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
Exceptions were analyzed for frequency and variability across providers.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N, % Agreement, Kappa (95% Confidence Interval)
Exceptions Reliability: 60, 100.0%, Kappa is non-calculable*

This measure demonstrates perfect reliability, as shown in the results from the above analysis.

The exception rate for this measure was 1.7%.
*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
This measure is not risk adjusted

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
This measure is not risk adjusted

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
Not Applicable

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
CMS Physician Quality Reporting Initiative:
Clinical Condition and Measure: #67
14,911 patients were reported on for the 2008 program, the most recent year for which data are available.

In 2009, the following was reported for this measure:
Eligible Professionals: 26,875
Professionals Reporting >=1 Valid QDC: 1,332
% Professionals Reporting >=1 Valid QDC: 4.96%
# Professionals Satisfactorily Reporting: 528
% Professionals Satisfactorily Reporting: 39.64%

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
CMS Physician Quality Reporting Initiative:
The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
CMS Physician Quality Reporting Initiative
This measure was used in the 2007-2011 CMS Physician Quality Reporting Initiative claims and registry options and group reporting option available in 2011.

There is a gap in care as shown by this 2008 data, the only year for which distribution by quartile/decile is available.

47.98% of patients reported on did not meet the measure.

10th percentile: 11.11%
25th percentile: 27.27%
50th percentile: 51.72%

75th percentile: 80.00%
90th percentile: 100.00%

The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 52.73, and indicates that 50% of physicians have performance on this measure ranging from 27.27% and 80.00%. A quarter of reporting physicians have performance on this measure which is greater than 80.00%, while a quarter have performance on this measure less than 27.27%.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
This measure has not been compared across data sources.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
This measure has not been compared across data sources.

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
This measure has not been compared across data sources.

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)*: We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one's ancestry) and language need (a rating of spoken English language proficiency of less than very well and one's preferred language for health-related encounters)."(2)

References:
(1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.

(2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

2.1-2.3 **Supplemental Testing Methodology Information:**

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

UPDATED MEASURE TESTING

Measure Number (*if previously endorsed*)**:** 377
Measure Title: Hematology: Myelodysplastic Syndrome (MDS) and Acute Leukemias: Baseline Cytogenetic Testing Performed on Bone Marrow
Date of Submission: 3/11/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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*).

Data (Registry)

The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

1.3. What are the dates of the data used in testing? The data are for the time period January 2014 through December 2014 and cover the entire United States.

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: 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*)

The total number of physicians reporting on this measure, via the registry reporting option, in 2014, is 151. Of those, 67 physicians had all of the required data elements and met the minimum number of quality reporting events (10) for a total of 1449 quality events. For this measure, 44.4 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 21.0 for the remaining 1,432 events. The range of quality reporting events for 67 physicians included is from 60 to 10. The average number of quality reporting events for the remaining 55.6 percent of physicians who aren't included is 0.26.

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

There were 1,432 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure and remained after exceptions were removed.

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.

The same data sample was used for reliability testing and exceptions analysis.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient-level socio-demographic (SDS) variables were not captured as part of the testing project for this measure.

2a2. RELIABILITY TESTING

***<u>Note</u>**: 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 of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

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*)

This measure has 0.68 reliability when evaluated at the minimum level of quality reporting events and 0.82 reliability at the average number of quality events.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Reliability at the minimum level of quality reporting events is moderate. Reliability at the average number of quality events is high.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

---

NQF Requirements of Inclusion of ICD-10 Codes:

NQF ICD-10-CM Requirement 1: Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.

NQF ICD-10-CM Requirement 2: See attachment in S.2b

NQF ICD-10-CM Requirement 3: The PCPI's ICD-10 conversion approach was used to identify ICD-10 codes for this measure. The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on

the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

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

The expert panel included 23 members. Panel members were comprised of experts from the ASH Committee on Quality. The list of expert panel members is as follows:

Gregory Abel, MD
Nathan Theodore Connell, MD, MPH
Mark A. Crowther, MD
Mary Cushman, MD, MSc
Adam Cuker, MD, MS
Reed Drews, MD
Joshua Field, MD
Nicola Goekbuget, MD
Lisa Kristine Hicks, MD,MSc,FRCPC
Vishal Kukreti, MD,FRCP,MSc
Jonathan D. Licht, MD
Wendy Lim, MD, MSc
Brea C. Lipe, MD
Gary H. Lyman, MD,MPH,FRCP
Navneet S. Majhail, MD, MS
Timothy McCavit, MD
Colleen Morton, MD, MS
Sarah H. O'Brien, MD
Menaka Pai, BSc,MD,FRCPC,MSc
Julie A. Panepinto, MD, MSPH
Anita Rajasekhar, MD
John J. Strouse, MD, PhD
William A. Wood Jr, MD, MPH

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

The results of the expert panel rating of the validity statement were as follows: N = 18; Mean rating = 4.61 and 94% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)
2 – 0 responses
3 – 1 responses (Neither Agree nor Disagree)
4 – 5 responses
5 – 12 responses (Strongly Agree)

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

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

Exceptions include:

- Documentation of medical reason(s) for not performing baseline cytogenetic testing (eg, no liquid bone marrow or fibrotic marrow)
- Documentation of patient reason(s) for not performing baseline cytogenetic testing (eg, at time of diagnosis receiving palliative care or not receiving treatment as defined above)
- Documentation of system reason(s) for not performing baseline cytogenetic testing (eg, patient previously treated by another physician at the time cytogenetic testing performed)

Exceptions were analyzed for frequency across providers.

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

Exceptions Analysis:

Amongst the 67 physicians with the minimum (10) number of quality reporting events, there were a total of 17 exceptions reported. The average number of exceptions per physician in this sample is 0.25 The overall exception rate is 1.2%.

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

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific or system reasons.

Without these being removed, the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives.

ASH recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. ASH also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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)

Not applicable

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

Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

Not applicable

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

Not applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

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

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

Measures of central tendency, variability, and dispersion were calculated.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
Based on the sample of 67 included physicians, the mean performance rate is 0.93, the median performance rate is 100%, and the mode is 1.00. The standard deviation is 0.13. The range of the performance rate is 0.74 , with a minimum rate of 0.26 and a maximum rate of 1.00. The interquartile range is 0.08 (.92 – 1).

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

The range of performance from 0.26 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note*: This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

This test was not performed for this measure.

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

This test was not performed for this measure.

2b6.3. 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*)

This test was not performed for this measure.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Data are not available to complete this testing.

2b7.2. 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*)

Data are not available to complete this testing.

2b7.3. 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*)

Data are not available to complete this testing.

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), 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)*

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.**

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.**

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA) or the American Society of Hematology (ASH). Neither ASH, nor the AMA, nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

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.

Planned	Current Use (for current use provide URL)
	Public Reporting PQRS http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/pqrs/index.html Professional Certification or Recognition Program ASH Myelodysplatic Syndromes PIM https://ashacademy.org/Product/index/1745

4a.1. For each CURRENT use, checked above, provide:

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

Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
Purpose: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). Eps satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013. Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented a phased approach to public reporting performance information on the Physician Compare Web site. CMS announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in late 2017.

The ASH MDS PIM is an MOC Practice Assessment activity intended for physicians seeking recertification in hematology and/or feedback on performance in this area of hematology. This is an ABIM-approved Practice Assessment activity, approved for 20 Practice Assessment points. The American Society of Hematology designates this PI CME activity for twenty (20) AMA PRA Category 1 Credits ™.

4a.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

4a.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

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

PQRS Experience Report

2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on Hematology: Myelodysplastic Syndrome (MDS) and Acute Leukemias: Baseline Cytogenetic Testing Performed on Bone Marrow were:

Average Performance Rate:
2010- 88.8%
2011- 94.6%
2012- 95.6%
2013- 87.0%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 5.7% of eligible professionals participating reported on MDS: Baseline Cytogenetic Testing. As a result, performance rates may not be nationally representative.

Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends.
Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

4b.2. **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.**
While the PCPI and ASH create measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1. Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences at this time, but we take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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

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 completely harmonized?

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.)
No related or competing measures

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):** American Society of Hematology
Co.2 **Point of Contact:** Robert, Plovnick, rplovnick@hematology.org, 202-629-5081-
Co.3 **Measure Developer if different from Measure Steward:** Physician Consortium for Performance Improvement
Co.4 **Point of Contact:** Caryn, Davidson, caryn.davidson@ama-assn.org, 312-464-4465-

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.

PCPI and ASH measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study are invited to participate as equal contributors to the measure development process. In addition, the PCPI and ASH strive to include on their work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Hematology Work Group
Steven L. Allen, MD (Co-Chair) (hematology/oncology)
William E. Golden, MD (Co-Chair) (internal medicine (IM))
Kenneth Adler, MD (hematology/IM)
Daniel Halevy, MD (nephrology)
Stuart Henochowicz, MD, MBA (IM)
Timothy Miley, MD (hematopathology)
David Morris, MD (radiation oncology)
John M. Rainey, MD (medical oncology)
Samuel M. Silver, MD, PhD (hematology/oncology)
Lawrence Solberg, Jr., MD, PhD (hematology/IM)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2007

Ad.3 **Month and Year of most recent revision:** 09, 2015

Ad.4 **What is your frequency for review/update of this measure?** Supporting guidelines, Specifications, and coding for this measure are reviewed annually.

Ad.5 **When is the next scheduled review/update for this measure?** 09, 2016

Ad.6 **Copyright statement:** The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA) or the American Society of Hematology (ASH). Neither ASH, nor the AMA, nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

The AMA and ASH encourage use of the Measures by other health care professionals, where appropriate.

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

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. ASH, the AMA, the PCPI and its members and former members of the AMA-PCPI disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

CPT® contained in the Measures specifications is copyright 2004-2015 American Medical Association. LOINC® copyright 2004-2015 Regenstrief Institute, Inc. SNOMED CLINICAL TERMS (SNOMED CT®) copyright 2004-2015 The International Health Terminology

Standards Development Organisation (IHTSDO). ICD-10 is copyright 2015 World Health Organization. All Rights Reserved. Ad.7 **Disclaimers:** Please see the copyright statement above in AD.6 for disclaimer information.
Ad.8 **Additional Information/Comments:**

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 #: 0378

Measure Title: Hematology: Myelodysplastic Syndrome (MDS): Documentation of Iron Stores in Patients Receiving Erythropoietin Therapy

Measure Steward: American Society of Hematology

Brief Description of Measure: Percentage of patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) who are receiving erythropoietin therapy with documentation of iron stores within 60 days prior to initiating erythropoietin therapy

Developer Rationale: In comparison with supportive care alone, patients receiving EPO with or without granulocyte colony-stimulating factor plus supportive care had improved erythroid responses, similar survival, and incidence of acute myeloid leukemia transformation (1). Treatment of anemia in MDS with EPO plus G-CSF was associated with significantly improved survival outcome in patients with no or low transfusion need, while not affecting the risk of leukemic transformation. Erythropoiesis-stimulating agents (ESAs: erythropoietin-alfa, darbepoietin) are a key component of the strategy for improving anemia and reducing dependence on red blood cell (RBC) transfusions. Clinical trial results indicate that approximately 40% of selected patients have a clinically meaningful hemoglobin response to ESAs, with a median two-year response. (2).To be effective, erythropoietin therapy requires that adequate iron stores be present due to iron's importance in red-blood-cell synthesis. By promoting the documentation of adequate iron stores in MDS patients requiring EPO therapy, the efficacy of the treatment will be enhanced (3).

1) Blood. 2009 Sep 17;114(12):2393-400. doi: 10.1182/blood-2009-03-211797. Epub 2009 Jun 29.
2) http://www.uptodate.com/contents/myelodysplastic-syndromes-mds-in-adults-beyond-the-basics
3) National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.

Numerator Statement: Patients with documentation of iron stores within 60 days prior to initiating erythropoietin therapy

Denominator Statement: All patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) who are receiving erythropoietin therapy

Denominator Exclusions: Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy

Measure Type: Process

Data Source: Electronic Clinical Data : Registry

Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Aug 09, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 Form
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 *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.

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:

- The evidence for this measure is based a clinical practice guideline from the National Comprehensive Cancer Network (NCCN) that states that iron repletion be verified before instituting Epo or darbepoetin therapy. Level of evidence: Category 2A
  - NCCN describes category 2A as based on "lower level evidence, there is uniform NCCN consensus that the intervention is appropriate."
  - The developer states that the quality of the body of evidence supporting the guideline recommendation was categorized as "lower- level evidence" which may include non-randomized trials; case series; or when other data are lacking, the clinical experience of expert physicians.
- The guideline does not a specify a timeframe that iron repletion be verified prior to Epo or darbepoetin therapy.

Updates: The developer provided updates to the guidelines – no changes were made to the evidence.

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:**

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as lower-level evidence (Box 6) → Low (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review.*
- *Is the SC aware of higher-level evidence to support this measure?*
- *Does the SC think there is a need to repeat the discussion and voting on 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.

The developer provided the following performance data from 1/1/2014 to 12/31/2014:

- **Registry Performance Rates**:
  - Mean: 54.58%
  - Minimum: 0.00%
  - Maximum: 100.00%
- **PQRS Average Performance Rates**:
  - 2010- 94.7%
  - 2011- 97.7%
  - 2012- 95.3%
  - 2013- 83.1% - Beginning in 2015, PQRS began imposing payment penalties for non-participants based on 2013 performance. For 2013, 6.5% of eligible professionals participating reported on this measure. As a result, performance rates may not be nationally representative.
- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).
- The developer provided additional data from the literature.

Disparities:

- The developer stated that federal reporting programs have not yet made disparities data available to analyze and report. Disparities data from the measure as specified is required for endorsement maintenance; this measure has been endorsed since 2008.
- The developer stated they are not aware of any literature outlining disparities for the documentation of iron stores in patients receiving erythropoietin therapy.

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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 to Support Measure Focus

Comments:

Process measure originally endorsed 2008 and most recently reviewed/endorsed 2012 has no new changes to evidence since last review. Remains consensus based support by ASCO, ASH and NCCN, level 2A, evidence low. Systematic review of evidence specific to measure exists. There is quality, quantity and consistency of evidence provided. Evidence is graded. SC member does not think there is a need to repeat discussion and voting on evidence. Reviewer does ask whether EPO level should also be included in measure as this is predictive of response to erythropoietin therapy for MDS low, intermediate risk patients.

I think that it applies directly and there is a relationship between the outcome and the action.

Yes.

1b. Performance Gap

Comments:

Yes, in addition to Registry performance rate with wide range and mean 54.58% and PQRS average performance rates, additional study published 2013 showed lack of concordance of NCCN guidelines use and community practice by longitudinal assessment within all MDS risk groups. Data not available for disparities in this area of healthcare.

No.

Yes; there is room for improvement.

A performance improvement plan is needed.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Registry

Specifications:

- This is a clinician-level measure
- The numerator includes patients with documentation of iron stores within 60 days prior to initiating erythropoietin therapy. Iron Stores include either: 1) bone marrow examination including iron stain OR 2) serum iron measurement including ferritin, serum iron and total iron-binding capacity (TIBC)
- The denominator includes all patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) who are receiving erythropoietin therapy. Erythropoietin therapy includes epoetin and darbepoetin.
    - The developer noted that regardless of when erythropoietin therapy was initiated, this measure is to be reported a minimum of once per reporting period for all MDS patients seen during the reporting period. The focus of the measure is that iron stores were *documented* within 60 days for each MDS patient receiving erythropoietin therapy regardless of when the therapy was initiated.
- Denominator exclusions include:
    - Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The calculation algorithm is provided.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability was conducted on a sample size from 2008 and 2009 that included 41 myelodysplastic syndrome (MDS) patient records from two hematology practice sites. Chart and data auditing occurred in 2010. The developer provided the percent agreement and kappa statistic (95% CI) for 41 charts:
    - Overall Reliability: 90.2%, 0.5470 (0.1578 – 0.9362)
    - Denominator Reliability: 100.0%
    - Numerator Reliability: 90.2%, 0.5470, (0.1578 – 0.9362)
    - Exceptions Reliability: 100.0%

Describe any updates to testing

- Reliability of the measure score was not presented in prior submission(s), reliability testing of the measure score has been conducted this review.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 2014 Registry data from PQRS. A total of 255 physicians reported on this measure in 2014. Of those, 28 physicians had 515 patient charts with all the required data elements and a **minimum of 10 quality reporting events**. The **average number of quality reporting events** (after exceptions were removed) was **18.4**.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability.

Results of reliability testing:

- Reliability at the at the minimum level of quality reporting events (10) was **0.88** and **0.93** at the average number of quality events (18.4).

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Appropriate method used/small sample size(Box 5) →Moderate reliability statistic and scope (Box 6) →Moderate

Questions for the Committee:

- *Is the test sample 28 physicians adequate to generalize for widespread implementation?*
- *Is it likely this measure can be consistently implemented?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing

- Additional empirical validity testing of the measure score has been conducted since the last review of this measure.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

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

Validity testing method

- The developer conducted new face validity testing with input from an expert panel including 23 members. The panel was comprised of experts from the ASH Committee on Quality.

Validity testing results:

- **89%** of the respondents either agreed or strongly agreed that this measure can accurately distinguish good and poor quality.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer notes several exclusions, as follows:
  - Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy
- The developer reported that there were a total of 97 exceptions reported amongst the 28 physicians with the minimum (10) number of quality reporting events. The average number of exceptions per physician was 3.5 and overall exception rate was 15.8%.
- Without the exclusions "the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives."
- The developer also stated that they recommend physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. ASH also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Questions for the Committee:

- *Are the results from the exclusion analysis a threat to validity?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer calculated measures of central tendency, variability, and dispersion. Based on a sample of 28 physicians:
  - Mean performance rate is 0.86
  - Median performance rate is 1.00, and the mode is 1.00
  - Standard deviation is 0.29
  - Range of the performance rate is 1.00 , with a minimum rate of 0.00 and a maximum rate of 1.00
  - Interquartile range is 0.02 (0.98 – 1.00)

Question for the Committee:

- *Does a sample size of 28 physicians demonstrate statistically significant and meaningful differences in quality across physicians?*

2b6. Comparability of data sources/methods:

- The developer stated test was not performed for this measure.

2b7. Missing Data

- The developer stated test was not performed for this measure.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Threats to validity mostly assessed (Box 2) →Empirical validity testing (Box 3)→ Face validity assessed (Box 4)→ Agreement measure can be used to distinguish quality (Box 5)→Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Inter rate reliability exists from prior review. Signal to noise ratio was high. Small physician sample size (n=28). Measure can be consistently implemented but should data element include EPO level?

The specs are appropriate.

**Face validity used with 89% agreement that this measure could distinguish good vs bad quality

Results indicate substantial agreement that the measure score could be used for quality assessment

MODERATE.**

Yes for validity.

2a2. Reliability Testing

Comments:

Yes.

2b2. Validity Testing

Comments:

Low sample number but overall low incidence of this disease making low sample size reasonable.

High percentage agreeing to the testing

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Exclusion analysis noted high total number of exception among the small number physicians could represent threat to validity.

No risk adjustments. sample relatively small.

Measuring iron stores adds value.

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 collection burden due to manual chart abstraction requirement.
- The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Data elements routinely generated by chart review or EHR in lab results.
Not used routinely.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:
- Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS). The developer stated that CMS announced that there are plans to make all PQRS individual EP level PQRS measures available for public reporting on Physician Compare in late 2017.
- ASH MDS PIM - an MOC Practice Assessment activity
- This measure has been endorsed since 2008 - per NQF criteria, performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results:
- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation:
- The developer reports no additional difficulties or unexpected findings or benefits, apart from those included throughout the submission form.

Potential harms: The developer reports no unintended consequence were noted.

Feedback:
- In 2012, prior to recommending the measure for maintenance endorsement, the Committee asked the developer to clarify the definition of "iron stores" in the numerator statement and to specify the time window for the denominator. The developer clarified the numerator and added the 60-day time window to the denominator for the documentation of iron stores prior to the initiation of erythropoietin therapy. The Committee agreed with the changes and recommended the measure for endorsement.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Current uses of measure not publicly reported but current use in an accountability program PQRS, ASH maintenance of certification practice assessment. CMS announced plans to make all PQRS individual EP level PQRS measures available for public reporting in late 2017.

Not publicly reported; yes, accountability in place.

Criterion 5: Related and Competing Measures

Related or competing measures

- No related or competing measures were identified.

Pre-meeting public and member comments

-

NQF #: 0378 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

Erythropoietin therapy offers a safer alternative than red blood cell transfusion for a wide range of anemic patients with MDS. To be effective erythropoietin requires that adequate iron stores be present due to iron's importance in red-blood-cell synthesis. Iron deficiency presents a major limitation to the efficacy of erythropoietin therapy.

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Both our measure and the NCCN guidelines state that iron repletion be verified before instituting Epo or darbepoetin therapy.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** The description of the evidence review in the guideline did not address the overall quantity of studies in the body of evidence. However NCCN guidelines for MDS reference 160 articles.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of the body of evidence supporting the guideline recommendation is summarized according to the NCCN categories of evidence and consensus as being based on "lower-level evidence". Lower-level evidence is later described as evidence that may include non-randomized trials; case series; or when other data are lacking, the clinical experience of expert physicians.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Although there is no explicit statement regarding the overall consistency of results across studies in the guidelines supporting the measure, the recommendation received uniform NCCN consensus that the intervention is appropriate.

In March 2007 and 2008, the FDA announced alerts and strengthened safety warnings for the use of Erythropoiesis-Stimulating Agents (ESAs).

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

To be effective erythropoietin requires that adequate iron stores be present due to iron's importance in red-blood-cell synthesis. Iron deficiency presents a major limitation to the efficacy of erythropoietin therapy.

1c.9 **Grading of Strength/Quality of the Body of Evidence.** Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** A panel of experts with members from each of the NCCN Member Institutions develops the NCCN Guidelines. Specialties that must be included on a particular panel are identified before that panel is convened but also evolve as the standard of care changes over time. This multidisciplinary representation varies from panel to panel. The NCCN Guidelines Panel Chairs are charged with ensuring that representatives of all treatment strategies are included. Many of the panels also include a patient representative, especially when issues of long-term care and patient preference are paramount in the panel´s considerations. NCCN publishes individual disclosures of potential conflicts of interest for panel members, NCCN Guidelines staff, and NCCN senior management. Relationships disclosed include research funding, participation in advisory groups, participation in speakers' bureaus, employment, and equity or patent ownership. Beginning in 2010, the NCCN Board of Directors has directed that panel members compensation from external sources be less than published thresholds. These thresholds are <= $20,000 from a single entity and <= $50,000 in aggregate from any source.

The ASCO Clinical Practice Guidelines Committee convened the ASCO/ASH Update Committee to lead the 2010 update. The Update Committee met via a series of teleconferences to review evidence collected from the systematic review and make revisions to the guideline recommendations as warranted.The guideline was reviewed and approved by the entire Update Committee, ASCO's Clinical Practice Guidelines Committee, ASH's Committee on Practice, ASH's Subcommittee on Quality of Care, the ASCO Board of Directors, and the ASH Executive Committee.

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** NCCN Categories of Evidence and Consensus Panel members identify the level of evidence supporting each recommendation. These categories are:

•Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

•Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

1c.13 **Grade Assigned to the Body of Evidence:** Category 2A

1c.14 **Summary of Controversy/Contradictory Evidence:** No controversy or contradictory evidence with regard to the importance of identifying documentation of iron stores in patients with MDS.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

1c.16 **Quote verbatim, <u>the specific guideline recommendation</u>** *(Including guideline # and/or page #)*:

Anemia related to MDS generally presents as a hypoproductive macrocytic anemia, often associated with suboptimal elevation of serum Epo levels. ~~To determined FAB subtype, iron status, and the level of ring sideroblasts, bone marrow aspiration with iron stain, biopsy, and cytogenetics should be examines~~. Bone marrow aspiration with iron stain, biopsy, and cytogenetics should be used to determine WHO subtype, iron status, and the level of ring sideroblasts. Patients should also be considered for HLA-DR15 typing as indicated above. Iron repletion needs to be verified before instituting Epo or darbepoetin therapy. (NCCN 2012)

2010 recommendation by American Society of Hematology: This recommendation remains the same as in 2007. Baseline and periodic monitoring of iron, total iron-binding capacity, transferrin saturation, or ferritin levels and instituting iron repletion when indicated may help to reduce the need for ESAs, maximize symptomatic improvement for patients, and determine the reason for failure to respond adequately to ESA therapy.

1c.17 **Clinical Practice Guideline Citation:** National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2012.

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes.

Version 1, 2016.

J. Douglas Rizzo, Melissa Brouwers, Patricia Hurley, Jerome Seidenfeld, Murat O. Arcasoy, Jerry L. Spivak, Charles L. Bennett, Julia Bohlius, Darren Evanchuk, Matthew J. Goode, Ann A. Jakubowski, David H. Regan and Mark R. Somerfield. Approved by the American Society of Clinical Oncology Board of Directors on July 7, 2010. Approved by the Executive Committee of the American Society of Hematology on July 14, 2010. Available here: http://www.hematology.org/Practice/Guidelines/2934.aspx

1c.18 National Guideline Clearinghouse or other URL: www.nccn.org

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: See 1c.10 above

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus Panel members identify the level of evidence supporting each recommendation. These categories are:

•Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

•Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

•Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

1c.23 Grade Assigned to the Recommendation: Category 2A

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: Moderate **1c.26** Quality: Moderate**1c.27** Consistency**:** Moderate

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0378_Evidence_form_FINAL.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
In comparison with supportive care alone, patients receiving EPO with or without granulocyte colony-stimulating factor plus supportive care had improved erythroid responses, similar survival, and incidence of acute myeloid leukemia transformation (1). Treatment of anemia in MDS with EPO plus G-CSF was associated with significantly improved survival outcome in patients with no or low transfusion need, while not affecting the risk of leukemic transformation. Erythropoiesis-stimulating agents (ESAs: erythropoietin-alfa, darbepoietin) are a key component of the strategy for improving anemia and reducing dependence on red blood cell (RBC) transfusions. Clinical trial results indicate that approximately 40% of selected patients have a clinically meaningful hemoglobin response to ESAs, with a median two-year response. (2).To be effective, erythropoietin therapy requires that adequate iron stores be present due to iron's importance in red-blood-cell synthesis. By promoting the documentation of adequate iron stores in MDS patients requiring EPO therapy, the efficacy of the treatment will be enhanced (3).

1) Blood. 2009 Sep 17;114(12):2393-400. doi: 10.1182/blood-2009-03-211797. Epub 2009 Jun 29.
2) http://www.uptodate.com/contents/myelodysplastic-syndromes-mds-in-adults-beyond-the-basics
3) National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014
Registry Performance Rate:
Mean: 54.58%
Minimum: 0.00%
Maximum: 100.00%

2013 PQRS Experience Report
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on Hematology: Myelodysplastic Syndrome (MDS): Documentation of Iron Stores were:

Average Performance Rate:
2010- 94.7%
2011- 97.7%
2012- 95.3%
2013- 83.1%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment

penalties for non-participants based on 2013 performance. For 2013, 6.5% of eligible professionals participating reported on this measure. As a result, performance rates may not be nationally representative.

Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

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.

A 2013 study examined ESA treatment patterns in a large, population-based sample of Medicare beneficiaries diagnosed with MDS between 2001–2005. Longitudinal analyses described not only whether patients received ESAs, but the patterns over time and the relationship to diagnostic evaluation and transfusion use. Using the NCCN guidelines as a standard for appropriate care, they observed a frequent lack of concordance between practice and guidelines. Patients were frequently not targeted for therapy based on risk status*, as evidenced by high rates of use across all risk groups, or on the likelihood of achieving response, as evidenced by frequent lack of measurement of serum EPO levels prior to ESA use.

Davidoff AJ, Weiss SR, Baer MR, Ke X, Hendrick F, Zeidan A, and Gore SD. Patterns of erythropoiesis-stimulating agent use among Medicare beneficiaries with myelodysplastic syndromes and consistency with clinical guidelines. Leuk Res. 2013 June ; 37(6): 675–680. doi:10.1016/j.leukres.2013.02.021.

*Risk status is determined by the level of anemia which is directly related to the iron stores.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

While this measure is included in federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

We are not aware of any publications/evidence outlining disparities for the documentation of iron stores in patients receiving erythropoietin therapy.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Severity of illness

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

In the general population, the incidence rate of MDS is approximately 4.8 per 100,000 people per year. However, MDS occurs more greatly among older people occurring in 29.6 per 100,000 people 70-79 years old, and 55.8 per 100,000 people older than 80. (1)

Approximately 80% of MDS patients experience symptomatic anemia. (2)

The 5-year survival rate is 29%. Bone marrow blast percentage, number of cytopenias, and cytogenetics represent major factors determining outcomes for patients with MDS in prognostic models such as the International Prognostic Scoring System (IPSS). In addition to karyotype and MDS subtype, transfusion dependence is a key factor in the WHO-based Prognostic Scoring System (WPSS). The MDS are a group of understudied hematologic disorders, and MDS may be the underlying condition affecting some elderly

patients with unexplained anemia. With the current demographic trend, increasing disease morbidity (both incidence and prevalence) is expected in the near future. (3)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1) National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Myelodysplastic syndromes. Version 1, 2016.

2) Davidoff AJ, Weiss SR, Baer MR, Ke X, Hendrick F, Zeidan A, and Gore SD. Patterns of erythropoiesis-stimulating agent use among Medicare beneficiaries with myelodysplastic syndromes and consistency with clinical guidelines. Leuk Res. 2013 June ; 37(6): 675–680. doi:10.1016/j.leukres.2013.02.021.

3) Ma X. Epidemiology of Myelodysplastic Syndromes. Am J Med. 2012 Jul; 125(7 Suppl): S2–S5. doi: 10.1016/j.amjmed.2012.04.014

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable. Not a PRO-PM.

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 subcriteria 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):*

Cancer, Cancer : Hematologic

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Safety, Safety : Medication Safety

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.)*

The measure specifications are included as an attachment with this submission. Additional measure details may be found at http://www.hematology.org/Clinicians/Guidelines-Quality/PQRS/503.aspx,

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:** NQF0378__I9toI10_conversion.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

Supporting guidelines and coding included in the measure are reviewed on an annual basis. However, this annual review has not resulted in any changes for this measure.

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)*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the

calculation algorithm.
Patients with documentation of iron stores within 60 days prior to initiating erythropoietin therapy

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
At least once during the 12 consecutive month measurement period

S.6. **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, 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.
Numerator Definition:
Documentation of Iron Stores – Includes either: 1) bone marrow examination including iron stain OR 2) serum iron measurement including ferritin, serum iron and total iron-binding capacity (TIBC)

For Registry:
Report the CPT Category II code: 3160F - Documentation of iron stores prior to initiating erythropoietin therapy

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 18 years and older with a diagnosis of myelodysplastic syndrome (MDS) who are receiving erythropoietin therapy

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Denominator Note:
This measure is to be reported a minimum of once per reporting period for all myelodysplastic syndrome (MDS) patients seen during the reporting period, regardless of when erythropoietin therapy is initiated; the quality action being measured is that iron stores were documented for each MDS patient receiving erythropoietin therapy within 60 days of starting erythropoietin therapy, regardless of how far back the erythropoietin therapy initiated.

Denominator Definition:
Erythropoietin Therapy – Includes the following medications: epoetin and darbepoetin for the purpose of this measure

For Registry:
Patients aged >= 18 years
AND
Diagnosis for MDS (ICD-9-CM) [reportable through 9/30/2015]: 238.72, 238.73, 238.74, 238.75
Diagnosis for MDS (ICD-10-CM) [reportable beginning 10/01/2015]: D46.0, D46.1, D46.20, D46.21, D46.22, D46.4, D46.9, D46.A, D46.B, D46.C, D46.Z
AND
Patient encounter during the reporting period (CPT): 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, 99241, 99242, 99243, 99244, 99245
AND
CPT Category II 4090F: Patient receiving erythropoietin therapy

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a

therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. For the measure Myelodysplastic Syndrome (MDS): Documentation of Iron Stores in Patients Receiving Erythropoietin Therapy, exceptions may include system reasons. Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Additional details by data source are as follows:
For Registry:
Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy - Append modifier to CPT Category II code: 3160F-3P

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment or risk stratification.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).
2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.

3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator

4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified for this measure: include system reasons. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable. The measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable. The measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not Applicable

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Clinician : Team

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **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. The measure is not a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Testing_Attachment_NQF_378_MDS_Documentation_of_Iron_Stores_Final.docx

NATIONAL QUALITY FORUM

NQF #: 0378 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
PCPI Testing Project
- Two hematology practice sites representing various types, locations and sizes were identified to participate in testing the measures
- Site A was a hematology group practice with eight physicians that cared for hematology patients. Site B was a large multi-specialty group clinic with 13 physicians that cared for hematology patients.
- Site A had a document retrieval system rather than a full-fledged EHR where data was scanned in and required searching. Site B had a fully functional EHR.
- Both sites were located in urban/suburban regions
- Hematology patient visit volume was 150 per day at site A and 120-150 per day at site B.
- Both sites were instructed to select 120 patient records (20 with acute leukemias and 35 for each of the following diagnoses: MDS, multiple myeloma and CLL).
- At site A the number of patients in practice in 2009 by specialty area was as follows:
  - Myelodysplastic Syndrome (MDS): 145 patients
- At site B the number of patients in practice in 2009 by specialty area was as follows:
  - Myelodysplastic Syndrome (MDS): 15 patients
- For this measure, the sample size included 41 abstracted patient charts. Site B did not have 30 patients in 2008 and 2009 so only 11 patients were included in the sample for this measure.
- The measurement period (data collected from patients seen) was between 1/1/2009 through 12/31/2009. Due to an inability to obtain the required number of patient records for acute leukemia and MDS during the specified measurement period, site B also included patients from 2008.
- Chart auditing was performed between 5/17/2010 and 7/15/2010
- Data auditing was performed between 8/2/2010 and 9/14/2010

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
Data abstracted from patient records were used to calculate inter-rater reliability for the measure.
41 MDS patient records were reviewed.

Data analysis included:
- Percent agreement
- Kappa statistic to adjust for chance agreement

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
N, % Agreement, Kappa (95% Confidence Interval)
Overall Reliability: 41, 90.2%, 0.5470 (0.1578 – 0.9362)
Denominator Reliability: 41, 100.0%, Kappa is non-calculable*
Numerator Reliability: 41, 90.2%, 0.5470, (0.1578 – 0.9362)
Exceptions Reliability: 41, 100.0%, Kappa is non-calculable*

This measure demonstrates moderately reliable, as shown in results from the above analysis.

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
Our measure recommends only that iron repletion be verified before Epo therapy and the NCCN evidence states that iron repletion be verified before instituting Epo or darbepoetintherapy.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Both our measure and the NCCN guidelines state that iron repletion be verified before instituting Epo or darbepoetintherapy.

The expert panel consisted of the following 10 members(with specialties listed):

Steven L. Allen, MD (Co-Chair) (hematology/oncology)
William E. Golden, MD (Co-Chair) (internal medicine (IM))
Kenneth Adler, MD (hematology/IM)
Daniel Halevy, MD (nephrology)
Stuart Henochowicz, MD, MBA (IM)
Timothy Miley, MD (hematopathology)
David Morris, MD (radiation oncology)
John M. Rainey, MD (medical oncology)
Samuel M. Silver, MD, PhD (hematology/oncology)
Lawrence Solberg, Jr., MD, PhD (hematology/IM)

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert Work Group and the measures adjusted as needed. Other external review groups (i.e. focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity has been quantitatively assessed for this measure. Specifically, the work group members were asked to empirically assess face validity of the measure. The work group/expert panel consists of 10 members, whose specialties include oncology, internal medicine, and clinical pathology.

Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

The scale 1-5, where 1=Disagree; 3=Neither Disagree nor Agree; 5=Agree.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
The results of the expert panel rating of the validity statement were as follows: N = 8; Mean rating = 4.75.

Percentage in the top two categories (4 and 5): 100%

Frequency Distribution of Ratings
1 - 0
2 - 0
3 - 0
4 - 2
5 – 6

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project

- 41 Myelodysplastic Syndrome patient records were reviewed for this measure.
- The measurement period (data collected from patients seen) was between 1/1/2009 through 12/31/2009.
- Chart auditing was performed between 5/17/2010 and 7/15/2010.
- Data auditing was performed between 8/2/2010 and 9/14/2010.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Exceptions were analyzed for frequency and variability across providers.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N, % Agreement, Kappa (95% Confidence Interval)
Exceptions Reliability: 41, 100.0%, Kappa is non-calculable*

This measure demonstrates perfect reliability, as shown in results from the above analysis.

The exception rate for this measure was 2.4%.

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
This measure is not risk adjusted.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
This measure is not risk adjusted.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
Not applicable.

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

CMS Physician Quality Reporting Initiative:
Clinical Condition and Measure: #68
11,494 patients were reported on for the 2008 program, the most recent year for which data are available.

In 2009 the following was reported for this measure:
Eligible Professionals: 21,607
Professionals Reporting >=1 Valid QDC: 1,235
% Professionals Reporting >=1 Valid QDC: 5.72%
# Professionals Satisfactorily Reporting: 452
% Professionals Satisfactorily Reporting: 36.60%

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
CMS Physician Quality Reporting Initiative:
The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
CMS Physician Quality Reporting Initiative
This measure was used in the 2007-2011 CMS Physician Quality Reporting Initiative claims and registry options and group reporting option available in 2011.

There is a gap in care as shown by this 2008 data, the only year for which distribution by quartile/decile is available.

58.00% of patients reported on did not meet the measure.

10th percentile: 0.00%
25th percentile: 6.91%
50th percentile: 30.22%
75th percentile: 66.67%
90th percentile: 97.44%

The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 59.76, and indicates that 50% of physicians have performance on this measure ranging from 6.91% and 66.67%. A quarter of reporting physicians have performance on this measure which is greater than 66.67%, while a quarter have performance on this measure less than 6.91%.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
This measure has not been compared across data sources.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**
This measure has not been compared across data sources

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**
This measure has not been compared across data sources.

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)***:** We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one's ancestry) and language need (a rating of spoken English language proficiency of less than very well and one's preferred language for health-related encounters)."(2)

References:
(1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.

(2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

2.1-2.3 **Supplemental Testing Methodology Information:**

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

UPDATED MEASURE TESTING

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 378
Measure Title: Hematology: Myelodysplastic Syndrome (MDS): Documentation of Iron Stores in Patients Receiving Erythropoietin Therapy
Date of Submission: 3/11/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For <u>all</u> measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For <u>outcome and resource use</u> measures**, section **2b4** also must be completed.
- If specified for **<u>multiple data sources/sets of specificaitons</u>** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to <u>all</u> questions as instructed with answers immediately following the question. All information on

testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.

- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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*).

Data (Registry)
The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

1.3. What are the dates of the data used in testing? The data are for the time period January 2014 through December 2014 and cover the entire United States.

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: 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*)

The total number of physicians reporting on this measure, via the registry reporting option, in 2014, is 255. Of those, 28 physicians had all of the required data elements and met the minimum number of quality reporting events (10), for a total of 612 quality events. For this measure, 11.0 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 18.4 for the remaining 515 events. The range of quality reporting events for 28 physicians included is from 50 to 10. The average number of quality reporting events for the remaining 89 percent of physicians who aren't included is 1.9.

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 were 515 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure and remained after exceptions were removed.

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.

The same data sample was used for reliability testing and exceptions analysis.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient-level socio-demographic (SDS) variables were not captured as part of the testing project for this measure.

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 of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

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*)

This measure has .88 reliability when evaluated at the minimum level of quality reporting events and .93 reliability at the average number of quality events.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., *what do the results mean and what are the norms for the test conducted?*)

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is very high.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

---

NQF Requirements of Inclusion of ICD-10 Codes:

NQF ICD-10-CM Requirement 1: Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.

NQF ICD-10-CM Requirement 2: See attachment in S.2b

NQF ICD-10-CM Requirement 3: The PCPI's ICD-10 conversion approach was used to identify ICD-10 codes for this measure. The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

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

The expert panel included 23 members. Panel members were comprised of experts from the ASH Committee on Quality. The list of expert panel members is as follows:

Gregory Abel, MD
Nathan Theodore Connell, MD, MPH
Mark A. Crowther, MD
Mary Cushman, MD, MSc
Adam Cuker, MD, MS
Reed Drews, MD
Joshua Field, MD
Nicola Goekbuget, MD
Lisa Kristine Hicks, MD,MSc,FRCPC
Vishal Kukreti, MD,FRCP,MSc
Jonathan D. Licht, MD
Wendy Lim, MD, MSc
Brea C. Lipe, MD
Gary H. Lyman, MD,MPH,FRCP
Navneet S. Majhail, MD, MS
Timothy McCavit, MD
Colleen Morton, MD, MS
Sarah H. O'Brien, MD
Menaka Pai, BSc,MD,FRCPC,MSc
Julie A. Panepinto, MD, MSPH
Anita Rajasekhar, MD
John J. Strouse, MD, PhD
William A. Wood Jr, MD, MPH

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

The results of the expert panel rating of the validity statement were as follows: N = 18; Mean rating = 4.44 and 89% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)
2 – 0 responses
3 – 2 responses (Neither Agree nor Disagree)
4 – 6 responses
5 – 10 responses (Strongly Agree)

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

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

Exceptions include:

Documentation of system reason(s) for not documenting iron stores prior to initiating erythropoietin therapy

Exceptions were analyzed for frequency across providers.

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

Exceptions Analysis:

Amongst the 28 physicians with the minimum (10) number of quality reporting events, there were a total of 97 exceptions reported. The average number of exceptions per physician in this sample is 3.5. The overall exception rate is 15.8%.

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

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific or system reasons.

Without these being removed, the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives.

ASH recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. ASH also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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)

Not applicable

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

Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

Not applicable

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

Not applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

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

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

Measures of central tendency, variability, and dispersion were calculated.

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

Based on the sample of 28 included physicians, the mean performance rate is 0.86, the median performance rate is 1.00, and the mode is 1.00. The standard deviation is .29. The range of the performance rate is 1.0, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is .02 (0.98-1.00).

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

The range of performance from 0.00 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

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

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

This test was not performed for this measure.

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

This test was not performed for this measure.

2b6.3. 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*)

This test was not performed for this measure.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Data are not available to complete this testing.

2b7.2. 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*)

Data are not available to complete this testing.

2b7.3. 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*)

Data are not available to complete this testing.

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)*
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.**

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.**

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA) or the American Society of Hematology (ASH). Neither ASH, nor the AMA, nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

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.

Planned	Current Use (for current use provide URL)
	Public Reporting PQRS http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/pqrs/index.html Professional Certification or Recognition Program ASH Myelodysplatic Syndromes PIM https://ashacademy.org/Product/index/1745

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor

- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
Purpose: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013.
Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented a phased approach to public reporting performance information on the Physician Compare Web site. CMS announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in late 2017.

The ASH MDS PIM is an MOC Practice Assessment activity intended for physicians seeking recertification in hematology and/or feedback on performance in this area of hematology. This is an ABIM-approved Practice Assessment activity, approved for 20 Practice Assessment points. The American Society of Hematology designates this PI CME activity for twenty (20) AMA PRA Category 1 Credits ™.

4a.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

4a.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

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

2013 PQRS Experience Report
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on Hematology: Myelodysplastic Syndrome (MDS): Documentation of Iron Stores were:

Average Performance Rate:
2010- 94.7%
2011- 97.7%
2012- 95.3%
2013- 83.1%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 6.5% of eligible professionals participating reported on this measure. As a result, performance rates may not be nationally representative.

Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends.

Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

4b.2. **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.**
While the PCPI and ASH create measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1. Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences at this time, but we take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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

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 completely harmonized?

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.)
No related or competing measures.

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):** American Society of Hematology
Co.2 **Point of Contact:** Robert, Plovnick, rplovnick@hematology.org, 202-629-5081-
Co.3 **Measure Developer if different from Measure Steward:** PCPI
Co.4 **Point of Contact:** Caryn, Davidson, caryn.davidson@ama-assn.org, 312-464-4465-

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.
PCPI and ASH measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study are invited to participate as equal contributors to the measure development process. In addition, the PCPI and ASH strive to include on their work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Hematology Work Group
Steven L. Allen, MD (Co-Chair) (hematology/oncology)
William E. Golden, MD (Co-Chair) (internal medicine (IM))
Kenneth Adler, MD (hematology/IM)
Daniel Halevy, MD (nephrology)
Stuart Henochowicz, MD, MBA (IM)
Timothy Miley, MD (hematopathology)
David Morris, MD (radiation oncology)
John M. Rainey, MD (medical oncology)
Samuel M. Silver, MD, PhD (hematology/oncology)
Lawrence Solberg, Jr., MD, PhD (hematology/IM)

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 09, 2015
Ad.4 **What is your frequency for review/update of this measure?** Supporting guidelines, Specifications, and coding for this measure are reviewed annually.
Ad.5 **When is the next scheduled review/update for this measure?** 09, 2016

Ad.6 **Copyright statement:** The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA) or the American Society of Hematology (ASH). Neither ASH, nor the AMA, nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

The AMA and ASH encourage use of the Measures by other health care professionals, where appropriate.

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

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. ASH, the AMA, the PCPI and its members and former members of the AMA-PCPI disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:** Please see the copyright statement above in AD.6 for disclaimer information.

Ad.8 **Additional Information/Comments:**

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 #: 0389
Measure Title: Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients
Measure Steward: PCPI
Brief Description of Measure: Percentage of patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy who did not have a bone scan performed at any time since diagnosis of prostate cancer
Developer Rationale: Multiple studies have indicated that a bone scan is not clinically necessary for staging prostate cancer in men with a low risk of recurrence and receiving primary therapy. For patients who are categorized as low-risk, bone scans are unlikely to identify their disease. Furthermore, bone scans are not necessary for low-risk patients who have no history or if the clinical examination suggests no bony involvement. Less than 1% of low-risk patients are at risk of metastatic disease.

While clinical practice guidelines do not recommend bone scans in low-risk prostate cancer patients, overuse is still common. An analysis of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it (1). The analysis also found that the use of bone scans in low-risk patients leads to an annual cost of $4 million dollars to Medicare. The overuse of bone scan imaging for low-risk prostate cancer patients is a concept included on the American Urological Association's list in the Choosing Wisely Initiative as a means to promote adherence to evidence-based imaging practices and to reduce health care dollars wasted (2). This measure is intended to promote adherence to evidence-based imaging practices, lessen the financial burden of unnecessary imaging, and ultimately to improve the quality of care for prostate cancer patients in the United States.

Citations:

1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

2. American Urological Association. A routine bone scan is unnecessary in men with low-risk prostate cancer. Choosing Wisely Initiative. Released February 21, 2013. Accessed February 25, 2016.

Numerator Statement: Patients who did not have a bone scan performed at any time since diagnosis of prostate cancer
Denominator Statement: All patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence, receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy
Denominator Exclusions: Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons)

Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician)

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Aug 09, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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:

- The developer provided a best practice statement, a clinical practice guideline, and a systematic review of the body of evidence to demonstrate the use of bone scans for low risk prostate cancer patients is not supported by the evidence, is extremely costly, and unnecessarily exposes patients to radiation.
- The Prostate-Specific Antigen Best Practice Statement: 2009 Update from American Urological Association (AUA) recommended:
  - Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL**. Level of evidence: No grade assigned**
- The National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011 recommended:
  - For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors or symptomatic disease**. Level of evidence: NCCN grade 2A** (2A is defined as the recommendation is based on **lower-level evidence** and there is uniform NCCN consensus)
- The systematic review was associated with the development of the best practice statement and clinical guideline; the developer cites the number of studies associated with each though the details of the Quality, Quantity, and Consistency of the evidence was not provided.

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 AUA Best Practice Statement and NCCN guideline were updated in 2013 and 2016, respectively. There were no changes to the recommendations since the previous submission.

- The developer provided new evidence: ACR Appropriateness Criteria. Prostate cancer- pretreatment detection, staging, and surveillance. American College of Radiology. 2012
    - ⋯only patients with a PSA ≥20 ng/ml (with any T stage or Gleason score), locally advanced disease (T3 or T4 with any PSA or Gleason score), or Gleason score ≥8 (with any PSA or T stage) should be considered for a radionuclide bone scan [91,99,101]. Patients with skeletal symptoms or advanced-stage disease should also be considered candidates for bone scans. p. 7. **Level of evidence based on the ACR 2012 Appropriateness Criteria:** 8 - Usually appropriate (Rating Scale: 1,2,3 Usually not appropriate; 4,5,6 May be appropriate; 7,8,9 Usually appropriate)

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → New evidence rated as 'usually appropriate'(Box 6) → Moderate (without QQC from SR, MODERATE is highest potential rating)

Questions for the Committee:

- *Is the Committee willing to accept the prior evaluation? The updated evidence supports the measure focus and has a stronger level of 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.

- The developer provided the following PQRS EHR, Registry, and Part B Claims data from January 1, 2014 – December 31, 2014:

	EHR Performance Rate	Registry Performance Rate
Mean	90.76%	90.24%
Minimum	50.0%	0.00%
Maximum	100.0%	100.00%

PQRS Experience Report

	Average Performance Rates
2010	71.60%
2011	90.50%
2012	92.50%
2013	88.50%

- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).
- The developer provided additional data from the literature on prostate cancer patients who received a bone scan although it was not recommend.

Disparities:

- The developer did not provide any data on disparities from the measure as specified – this is encouraged for endorsement maintenance.
- The developer stated that while this measure is included in federal reporting programs, those programs have not yet made disparities data available to analyze and report.
- The developer provided a summary of data from the literature that compares the incidence, prevalence, and

death rates between African American men and white men due to prostate cancer.

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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**

Rationale: Insufficient information provide to determine if a performance gap exists.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

This is a process measure designed to prevent unnecessary burden of cost and radiation to asymptomatic men with low and very low risk prostate cancer. Using systematic review and guideline statements from NCCN, AUA and (new this submission) ACR, the developers demonstrate that this population has <1% of harboring bone metastasis. ACR provided their highest rating (usually appropriate) to this recommendation. Thus, the updated evidence has a stronger level of evidence leading to a Moderate rating.

There continues to be a strong level of evidence for this measure, with support from the AUA, ACR, and NCCN as well as a systematic review.

1b. Performance Gap

Comments:

**This category was rated as INSUFFICIENT. The developers cite literature from 2004-2007 demonstrating that 43% of patients wtih low or very low risk prostate cancer have bone scans despite recommendation against, translating to $4 million annually. Data provided from registry and claims data for the year 2014 show compliance in the range of 90%, ranging from 0 to 100%. This has increased from average performance of 72% in 2010.

Disparity data are limited and have not yet been provided by federal programs. They cite literature showing higher morbidity and mortality of prostate cancer in African Americans. Another citation suggests that imaging overuse is associated with non-white race, education and income measures and region (highest in northeast).**

No data provided except national numbers of men at risk and estimates of cost of bone scan and down stream imaging in this population: 4 million. The developers indicate that the measure is a high priority to due high resource use.

This is an area of question. This measure has been endorsed since 2008 at this time and appears to have performance rates in the high 80s to low 90s. This was relatively stable over time as well. Disparities data is not available. The question is therefore whether a true gap exists.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): electronic clinical data, registry data. This is not an e-Measure.

Specifications:

- This is a clinician-level measure.
- The numerator includes patients who did not have a bone scan performed at any time since diagnosis of prostate cancer.
- The denominator includes all patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence, receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate,

OR radical prostatectomy, OR cryotherapy.
    - The developer provided risk strata definitions for very low, low, intermediate, high, or very high and external beam radiotherapy.
- Denominator exclusions include:
    - Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons), and
    - Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician).
- ICD 9, ICD 10, and CPT codes are included.
- The calculation algorithm is provided.
- Missing data either delete a case from the denominator or represent a quality failure.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability was conducted on a 94 patient records from 2010. Chart and data auditing occurred in 2011. The developer provided the percent agreement and kappa statistic (95% CI):
    - Overall Reliability: N=94, 100%
    - Denominator Reliability: N=94, 100%
    - Numerator Reliability: N=94, 100%
    - Exceptions Reliability: N=94, 100%
    - Kappa Statistics could not be calculated because of complete agreement. Confidence intervals could not be calculated because to do so would involve dividing by zero which cannot be done.

Describe any updates to testing: Reliability testing of the measure score – see below

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset included 2014 Registry data from PQRS provided by CMS. A total of 131 physicians reported on this measure in 2014. Of those, 24 (18.3%) physicians had 1,113 patient charts with all the required data elements and a **minimum of 10 quality reporting events**. The **average number of quality reporting events** (after exceptions were removed) was **46.0**.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability.

Results of reliability testing:

- Reliability at the at the minimum level of quality reporting events (10) was **0.84** and **0.96** at the average number

of quality events (46.0).

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Appropriate method used (Box 5) →High/moderate reliability statistic and scope (Box 6) →Moderate

Questions for the Committee:

- *Is a test sample of 24 physicians adequate to generalize for widespread implementation?*
- *Is it likely this measure can be consistently implemented?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing: Additional empirical validity testing of the measure score has been conducted since the last review of this measure.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- Face validity was assessed using a panel of 17 experts from representation from the PCPI Measures Advisory Committee.

Validity testing results:

- **80%** (10) of the respondents either agreed or strongly agreed that this measure can accurately distinguish good and poor quality.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The exclusions include:
  - Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons), and
  - Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician).
- The developer noted that the 24 physicians with the minimum (10) number of quality reporting events reported 183 exceptions from January 2014-December 2014.
  - The average number of exceptions per physician in the sample (131) was 7.6.
  - The overall exception rate was 14.1%.
- The developer stated that this methodology does not require external reporting of more detailed exception data but does recommend that physicians document the specific reasons for the exception in patients' medical records.

Questions for the Committee:

- *Are the results from the exclusions analysis a threat to validity? Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the seemingly high number of exclusions reasonable?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer calculated measures of central tendency, variability, and dispersion. Based on a sample of 24 physicians:
  - Mean performance rate is 0.89
  - Median performance rate is 1.0, and the mode is 1.00 – *this data provided by the developer may an error*
  - Standard deviation is 0.20
  - Range of the performance rate is 0.49, with a minimum rate of 0.51 and a maximum rate of 1.00
  - Interquartile range is 0.15 (0.88 – 1.00)

Question for the Committee:

- *Based on a sample of 24 physicians, does this measure identify stistically and clinically meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Measure is not specified for more than one data source; comparability of data sources is not needed.

2b7. Missing Data

- The developer addressed how to handle missing data in S22; If data required to determine if a denominator eligible patient qualifies for the numerator (or has a valid exclusion/exception) are missing, this case would represent a quality failure.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Threats to validity mostly assessed (Box 2) →Empirical validity testing (Box 3)→ Face validity assessed (Box 4)→ Agreement measure can be used to distinguish quality (Box 5)→Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**Numerator--clearly defined: no bone scan since time of diagnosis of prostate cancer.
Denominator--men receiving local therapy of any kind for low or very low risk prostate cancer as defined by updated risk strata reported by NCCN (Gleason 6, PSA<10, stage T1 to 2a). As stated, this measure will only capture patients receiving active intervention. The increasing number of men opting for "active surveillance"--who presumably would also not be candidates for bone scan--will not be captured by this measure.**
The measures appears to be clearly defined and able to be consistently implemented. The algorithm/specifications are clear.
Measure specifications are consistent with the evidence --these populations are very unlikely to have bone positive disease.
The face validity was agreed upon by the experts 80% of the time. The score can be used to assess quality and is an **indicator thereof.**

2a2. Reliability Testing

Comments:

Measure score. 2014 registry data from PQRS had 131 physicians who reported, with 24 meeting required minimum 10 reporting events. Of these 24, reliability score was 0.84 at minimum (10) events and 0.96 at the average number of events (46). The reporting is reliable in this small subset.
It is a bit unclear to me why only 18.3% of physicians were able to be included. Was that because elements were missing or that the required number of reporting events was infrequent? If the prior, could be problematic.

2b2. Validity Testing

Comments:

**Only face validity was performed. An expert panel of 17 urologists, methodologists, clinical and radiation oncologists, pathologists, primary care physicians and consumer representatives from the PCPI Measures Advisory Committee were asked to rank the measure on a 5-point scale as to whether they agreed it could distinguish good from poor quality. Of TEN respondents, 2 disagreed and 8 agreed or strongly agreed. Not an overwhelming consensus.
I do agree that the score is an indicator of quality as it pertains to appropriate use of healthcare resources and limitation of unnecessary radiation exposure.**
An exclusion rate of 14% appears reasonable in clinical practice.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias

Comments:

**Exclusions include scan physician judgement of cause (pain, salvage or other medical reason) OR scan ordered by another physician. This reason must be specifically documented in medical record. Overall exception rate was 14.1 with average of 7.6 per physician. Mean performance rate was 0.89 (range 0.51-1.00). Rated: MODERATE
Any cases where the cause of exclusion is missing will be counted as failure to meet the measure. Note is made that patients found to have a POSITIVE bone scan will remain in the denominator even though they no longer meet LR or VLR risk strata.**
While there is a method outlined to address missing data, it is unclear how often this occurs. Key elements meant to risk stratify a patient would be a threat to the ability to implement 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.

- This measure is based on clinical registry data and all data elements are available in electronic sources.
- The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:

Based on clinical registry data available in electronic sources. The primary elements are routinely generated, though exclusion data may be less reliably extracted.

All are generated during care and used to stratify patients. Whether some may be missing at the time of the scanning decision is unclear, so the analysis would have to be made based on the data available at that time.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details:

- Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS). The developer stated that CMS announced that there are plans to make all PQRS individual EP level PQRS measures available for public reporting on Physician Compare in late 2017.
- This measure has been endorsed since 2008 - per NQF criteria, performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results:

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation:

- The developer has not identified any additional difficulties or unexpected findings or benefits during the implementation of this measure.

Potential harms:

- The developer stated that they are not aware of any unintended consequences at this time, but take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

Feedback :

- In 2012 NQF evaluation of this measure, public and member comments received included:
  - Commenters indicated that the Steering Committee should consider clarifying 'low risk' status for the measure population and that classification for measurement purposes should be based on staging information available at the time of decision making regarding whether or not to order a bone scan.
  - Commenters believed that the measure should clearly articulate that even those patients with a positive bone scan remain in the denominator of this measure, even though the bone scan ultimately demonstrates that they are not actually low risk.
  - Comments reflected questions on the measure specifications, specifically:
    - It is unclear how treatment interplays with this measure.
    - The numerator captures patients who did not have a 'bone scan performed prior to initiation of treatment nor at any time since diagnosis.
    - Patient eligibility for the denominator should be based on criteria known before the decision to deliver the service (the bone scan) is considered.
    - Exclusion criteria (i.e. treatment planned for future, patient preference, vulnerable health status, and poor access to care)
  - Several commenters supported this measure.
- The measure developer's response was:
  - The AUA/AMA-PCPI Prostate Cancer Work Group appreciates your comment. The Work Group will consider your feedback about the risk stratification, when the measure undergoes formal review and maintenance, according to the AMA-PCPI measure development/maintenance methodology, in the future. Additionally, the measure contains a medical exception, which allows physicians to use clinical judgment in order to have a bone scan performed on those low-risk prostate cancer patients who have a medical reason documented.
  - The denominator was constructed so any patient that has already been stratified as a low risk patient and is being treated according to the low risk strata would be captured in the measure. The measure is aiming to reduce the use of bone scans that are clinically unnecessary, in low risk patients who generally have no indication for imaging studies. Additionally, the measure contains a medical exception, which allows physicians to use clinical judgment in order to n performed on those low-risk prostate cancer patients who have a medical reason documented.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Will be publicly reported in late 2017. currently in use in "at least 1" accountability program. No unintended consequences have been encountered since initial approval. This measure should continue to identify areas where overuse of bone scan is practiced and help to minimize this burden.

The measure can be used to address a quality gap if unnecessary testing is occurring. The main consequence of decreasing testing rates would be missing metastatic disease, but this is likely to be quite rare per the evidence.

Criterion 5: Related and Competing Measures

Related or competing measures:

- 0390 : Prostate Cancer: Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients
- 1853 : Radical Prostatectomy Pathology Reporting

Harmonization:

- 0390 and 1853 measure different target populations and addresses different aspects of prostate cancer care.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0389 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The process of identifying the patient´s risk strata prior to ordering any imaging studies is related to improved outcomes, including cost reduction and reduction of radiation exposure.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The evidence directly supports the specified measure. The measure specifically identifies the risk strata for whom bone scans are inappropriate. The guideline and best practice statement do not recommend bone scans for patients included in the low or very low risk strata.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: The AUA best practice statement references a systematic review of 23 studies, examining the utility of bone scan.

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

The description of the evidence review in the NCCN guideline did not address the overall quantity of studies in the body of evidence. However, 223 articles are cited in NCCN´s prostate cancer guideline´s reference section.

AUA 2013 Guideline:
The guideline cites that 23 studies were reviewed to develop the recommendation statement.

NCCN 2016 Guideline:
Information regarding the total number of studies and type of study designs included in the body of evidence is not available. However, the guideline cites 1 observational study in support of the recommendation statement.

ACR 2012 Appropriateness Criteria:
The guideline cites 6 observational diagnostic studies and 3 reviews/other diagnostic studies in support of the recommendation statement.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the*

evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events): The systematic review of the literature (cited within the AUA best practice statement) states the following:

Studies were eligible only if newly diagnosed PC cases with no previous management were included. Studies were excluded if details regarding the patient population or results were significantly lacking.

The findings of this study are based on data pooled primarily from retrospective series. It is possible that inherent biases occurred in reporting. For example although not always reported, it may have been that in some studies those cases with a positive CT did not proceed to surgery. Results from these cases may not have been reported, thereby lowering the apparent detection rate in the reported population.

As with all studies of this nature, there are limitations to the findings. Unfortunately, not all series reported results based on the pooled groupings of PSA, tumor stage and Gleason score used herein. In addition, not all studies graded disease using Gleason score. As a result, inclusion of data from all cases was not justified. Fortunately, large patient numbers were remaining to allow for small confidence intervals around estimates.

Because of the nature of this study, it is not possible to make recommendations based on combinations of prognostic factors. For example bone scanning detected metastases in 6.4% of patients with localized disease but it is not possible to tell what proportionwere at risk by virtue of increased PSA or Gleason score, for which scanning would have been indicated. Presumably, some of those patients with positive bone scans would have been at risk from either of these factors. Therefore, it could be argued that the true risk for patients with low PSA, low Gleason score and localized disease is less than those numbers reported here. Also, most of these studies were published in the 1990s and contained results for patients seen before the widespread use of PSA screening. Therefore, no distinction can be made in patients with organ confined disease between those with palpable and nonpalpable tumors. Again, it could be argued that due to stage migration within this group, numbers reported here are higher than the true risk.

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

The quality of the body of evidence supporting the NCCN guideline recommendation is summarized according to the NCCN categories of evidence and consensus as being based on "lower-level evidence."

AUA 2013 Guideline:
The quality of body of evidence was not included.

NCCN 2016 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence, however, the guideline is a Category 2A: "Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate."

ACR Appropriateness Criteria:
The quality of body of evidence was not included.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The systematic review referenced by the AUA best practice statement does not contain information about consistency of results across studies.

Although there is no explicit statement regarding the overall consistency of results across studies in the NCCN guideline, the recommendation received uniform NCCN consensus that the recommendation is appropriate.

AUA 2013 Guideline:
The consistency of results across studies was not reviewed.

NCCN 2016 Guideline:
The guideline does not provide the consistency of results across studies, however, the recommendation received uniform NCCN consensus that the intervention is appropriate.

ACR Appropriateness Criteria:
The consistency of results across studies was not reviewed.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**
Overuse of bone scans among prostate cancer patients is extremely costly and unnecessarily exposes patients to radiation. The use of bone scans for low risk prostate cancer patients is not supported by evidence.

AUA 2013 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence. However, they do include the following summary information regarding the benefits of not performing a bone scan on low-risk prostate cancer patients:
"The authors concluded that low-risk patients are unlikely to have disease identified by bone scan. Accordingly, bone scans are generally not necessary in patients with newly diagnosed prostate cancer who have a PSA < 20 ng/mL unless the history or clinical examination suggests bony involvement."

While no harms were mentioned, it is expected that the harms of using a bone scan on low-risk patients includes unnecessary exposure to radiation which can have potential harms such as radiation burns, adverse reactions to contrast media, and radiation-induced malignancy.

NCCN 2016 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence. However, the guideline does state "Patients with low-and intermediate-risk and low postoperative serum PSA levels have a very low risk of positive bone scans or CT scans. In a series of 414 bone scans performed in 230 men with biochemical recurrence after radical prostatectomy, the rate of a positive scan for men with PSA >10 ng/mL was only 4%."

While the guidelines did not describe how the harms studied affected net benefits, they did state that the risks of imaging include adverse reaction to contrast media, false-positive scans, and over-detection. Additional risks of imaging include radiation burns, cataracts, radiation-induced malignancy, and adverse reactions to contrast material delivered by intravenous, oral, or rectal routes.

ACR 2012 Appropriateness Criteria:
The guideline does not include an overall estimate of benefit from the body of evidence. However, they do include the following summary information regarding the benefits of not performing a bone scan on low-risk prostate cancer patients:
"Work by Oesterling and others has shown that in patients with low PSA level (<10 ng/ml) who have no pain, the yield of a staging bone scan is too low to warrant its routine use. In their experience, no patient with a PSA ≤ 10 ng/ml had a positive bone scan and only one patient in 300 with a PSA level ≤ 20 ng/ml had a positive radionuclide scan"

While no harms were mentioned, it is expected that the harms of using a bone scan on low-risk patients includes unnecessary exposure to radiation which can have potential harms such as radiation burns, adverse reactions to contrast media, and radiation-induced malignancy.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

AUA 2013 Guideline:
No grade has been assigned for the quality of evidence.

NCCN 2016 Guideline:
Yes

ACR 2012 Appropriateness Criteria:
Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: NCCN Prostate Cancer Panel

Andrew J. Armstrong, MD, ScM
Robert R. Bahnson, MD
Barry Boston, MD
J. Erik Busby, MD
Anthony Victor D´Amico, MD, PhD
James A. Eastham, MD
Charles A. Enke, MD
Thomas A. Farrington
Lauren Gallagher, RPh, PhD
Kristina M. Gregory, RN, MSN, OCN
Celestia S. Higano, MD, FACP
Maria Ho, PhD
Eric Mark Horwitz, MD
Philip W. Kantoff, MD
Mark H. Kawachi, MD
Michael Kuettel, MD, MBA, PhD
Richard J. Lee, MD, PhD
Gary R. MacVicar, MD
Arnold W. Malcolm, MD, FACR
Joan S. McClure, MS
David Miller, MD, MPH
James L. Mohler, MD
Elizabeth R. Plimack, MD, MS
Julio M. Pow-Sang, MD
Mack Roach, MD
Eric Rohren, MD, PhD
Stan Rosenfeld
Dorothy Shead, MS
Sandy Srinivas, MD
Seth A. Strope, MD, MPH
Jonathan Tward, MD, PhD
Przemyslaw Twardowski, MD
Patrick C. Walsh, MD

The NCCN Guidelines are updated at least annually in an evidence-based process integrated with the expert judgment of multidisciplinary panels of expert physicians from NCCN Member Institutions. NCCN depends on the NCCN Guidelines Panel Members to reach decisions objectively, without being influenced or appearing to be influenced by conflicting interests.

All panel member disclosures are available at www.nccn.org.

NCCN 2016 Guideline Prostate Cancer Panel:
James Mohler, MD
Andrew Armstrong, MD
Robert Bahnson, MD
Anthony Victor D'Amico, MD PhD
Brian Davis, MD, PhD
James Eastham, MD
Charles Enke, MD
Thomas Farrington,
Celestia Higano, MD
Eric Horwitz, MD
Michael Hurwitz, MD, PhD
Christoper Kane, MD
Mark Kawachi, MD

Michael Kuettel, MD, MBA, PhD
Richard Lee, MD, PhD
Joshua Meeks, MD, PhD
Daivd Penson, MD, MPH
Elizabeth Plimack, MD, MS
Julio Pow-Sang, MD
David raben, MD
Sylvia Richey, MD
March Roach, III, MD
Stan Rosenfeld
Edward Schaeffer, MD, PhD
Ted Skolarus, MD
Eric Small, MD
Guru Sonpavde, MD
Sandy Srinivas, MD
Seth Strope, MD, MPH
Johnathon Tward, MD, PhD

All panel member disclosures are available at www.nccn.org.

ACR 2012 Appropriateness Criteria: Expert Panels on Urologic Imaging and Radiation Oncology–Prostate:
Steven C. Eberhardt, MD
Scott Carter, MD
David D. Casalino, MD
Gregory Merrick, MD
Steven J. Frank, MD
Alexander R. Gottschalk, MD, PhD
John R. Leyendecker, MD
Paul L. Nguyen, MD
Aytekin Oto, MD
Christopher Porter, MD
Erick M. Remer, MD
Seth A. Rosenthal, MD

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guideline:
NCCN Categories of Evidence and Consensus
Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

ACR 2012 Appropriateness Criteria:
Study Quality Category Definitions
- Category 1 The study is well-designed and accounts for common biases.
- Category 2 The study is moderately well-designed and accounts for most common biases.
- Category 3 There are important study design limitations.
- Category 4 The study is not useful as primary evidence. The article may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus. For example:

a) the study does not meet the criteria for or is not a hypothesis-based clinical study (e.g., a book chapter or case report or case series description);
b) the study my synthesize and draw conclusions about several studies such as a literature review article or book chapter but is not primary evidence;
c) the study is an expert opinion or consensus document.

1c.13 Grade Assigned to the Body of Evidence: No grade for AUA best practice statement, NCCN grade 2A

AUA 2013 Guideline:
No grade has been assigned to the body of evidence.

NCCN 2016 Guideline:
Level of evidence assigned: Category 2A

ACR 2012 Appropriateness Criteria:
The guideline includes 5 observational diagnostic studies with a study quality of 3 and 1 observational diagnostic study with a study quality of 2. The guideline also includes 3 reviews/other diagnostic studies with a study quality of 4.

1c.14 Summary of Controversy/Contradictory Evidence: No contradictory evidence has been identified.

AUA 2013 Guideline:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

NCCN 2016 Guideline:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

ACR 2012 Appropriateness Criteria:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
A radionuclide bone scan is traditionally the first examination obtained. If the bone scan is positive for metastatic disease, no further imaging studies are
necessary. If it is inconclusive, further imaging studies are performed, including conventional radiographs, MRI, or computed tomography (CT). However, the level of posttreatment PSA that should prompt a bone scan is uncertain. In a study of patients with biochemical failure following radical prostatectomy, the probability of a positive bone scan was less than 5% with PSA levels between 40-45 ng/ml. In another study, bone scan was limited until PSA rose above 30-40 ng/ml. Men with a PSADT of <6 months after radical prostatectomy were at increased risk of a positive bone scan (26% vs 3%) or
positive CT (24% vs 0%) compared to those with longer PSADT. Kane et al reported that most patients with a positive bone scan had a high PSA level (mean of 61.3 ng/ml) and a high PSA velocity (>0.5 ng/ml/month).

American College of Radiology. ACR Appropriateness Criteria. Post-treatment Follow-up of Prostate Cancer. 2011. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/app_criteria/pdf/ExpertPanelonUrologicImaging/PostTreatmentFollo

wUpofProstateCancerDoc10.aspx

The results of a retrospective review demonstrate extensive overuse of bone scan imaging among VA patients with low-risk prostate cancer.Overall, the rate of bone scan imaging among men with low-risk features was 25% with no positive findings.

Palvolgyl R, Daskivich TJ, Chamie K, Kwan L, Litwin MS. Bone Scan Overuse in Staging of Prostate Cancer: An Analysis of a Veterans Affairs Cohort.

Citation for the systematic review of literature, referenced in sections 1c.5 and 1c.6 is below:

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
1. Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL.

2. For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors or symptomatic disease.

AUA 2013 Guideline:
Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL. p. 4

NCCN 2016 Guideline:
For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors; or 4) symptomatic disease. (Category 2A) p. 64

ACR 2012 Appropriateness Criteria:

Clinical Condition: **Prostate Cancer — Pretreatment Detection, Staging, and Surveillance**

Variant 3: **Prostate cancer diagnosed on biopsy, patient at high risk for locally advanced disease and metastases (AJCC Groups III and IV). Example: PSA ≥20 or Gleason 8-10 or clinical stage T2c or higher.**

Radiologic Procedure	Rating	Comments	RRL*
MRI pelvis without and with contrast	8	Should include dynamic contrast-enhanced (DCE) technique. See statement regarding contrast in text under "Anticipated Exceptions."	O
Tc-99m bone scan whole body	8		☢☢☢
CT abdomen and pelvis with contrast	7		☢☢☢☢
MRI pelvis without contrast	6		O
CT abdomen and pelvis without contrast	6	If contrast contraindicated.	☢☢☢☢
X-ray area of interest	4	Appropriate if bone scan or symptoms suggest possible involvement.	Varies
FDG-PET/CT whole body	4		☢☢☢☢
CT abdomen and pelvis without and with contrast	2		☢☢☢☢
In-111 capromab pendetide scan	2		☢☢☢☢
Rating Scale: 1,2,3 Usually not appropriate; 4,5,6 May be appropriate; 7,8,9 Usually appropriate			***Relative Radiation Level**

…only patients with a PSA ≥20 ng/ml (with any T stage or Gleason score), locally advanced disease (T3 or T4 with any PSA or

Gleason score), or Gleason score ≥8 (with any PSA or T stage) should be considered for a radionuclide bone scan [91,99,101]. Patients with skeletal symptoms or advanced-stage disease should also be considered candidates for bone scans. p. 7

1c.17 Clinical Practice Guideline Citation: 1. Prostate-Specific Antigen Best Practice Statement: 2009 Update from American Urological Association, American Urological Association Education and Research, Inc. Available at:http://www.auanet.org/content/guidelines-and-qualitycare/clinical-guidelines/main-reports/psa09.pdf.

2. National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011. Available at www.nccn.org

AUA 2013 Guideline:
Carroll P, Albertsen PC, Greene K, et al. American Urological Association Education and Research, Inc. PSA testing for the pretreatment staging and posttreatment management of prostate cancer: 2013 Revision of 2009 Best Practice Statement. Linthicum, MD: American Urological Association Education and Research, Inc. 2013. Available at: https://www.auanet.org/common/pdf/education/clinical-guidance/Prostate-Specific-Antigen.pdf

NCCN 2016 Guideline:
National Comprehensive Cancer Network (NCCN). Clinical practice guidelines in oncology: prostate cancer. Version 2.2016. Available at www.nccn.org

ACR 2012 Appropriateness Criteria
Eberhardt SC, Carter S, Casalino D, et al. ACR Appropriateness Criteria. Prostate cancer- pretreatment detection, staging, and surveillance. American College of Radiology. 2012. Available at: https://acsearch.acr.org/list

1c.18 National Guideline Clearinghouse or other URL: http://www.auanet.org/content/media/psa09.pdf and www.nccn.org

AUA 2013 Guideline:
Available at http://www.auanet.org/education/aua-guidelines.cfm

NCCN 2016 Guideline:
Available at NCCN.org

ACR 2012 Appropriateness Criteria
Available at: https://acsearch.acr.org/list

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

AUA 2013 Guideline:
No

NCCN 2016 Guideline:
Yes

ACR 2012 Appropriateness Criteria
Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: NCCN Prostate Cancer Panel is listed in section 1c.10

NCCN 2016 Guideline:
NCCN Prostate Cancer Panel is listed in section 1c.10

ACR 2012 Appropriateness Criteria
Expert Panels on Urologic Imaging and Radiation Oncology–Prostate is listed in section 1c.10

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guideline:
NCCN Categories of Evidence and Consensus
Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

ACR 2012 Appropriateness Criteria:
ACR Appropriateness Criteria Methodology
The ACR AC methodology is based on the RAND Appropriateness Method2. The appropriateness ratings for each of the procedures or treatments included in the AC topics are determined using a modified Delphi method. A series of surveys are conducted to elicit each panelist's expert interpretation of the evidence, based on the available data, regarding the appropriateness of an imaging or therapeutic procedure for a specific clinical scenario. The expert panel members review the evidence presented and assess the risks or harms of doing the procedure balanced with the benefits of performing the procedure. The direct or indirect costs of a procedure are not considered as a risk or harm when determining appropriateness. When the evidence for a specific topic and variant is uncertain or incomplete, expert opinion may supplement the available evidence or may be the sole source for assessing the appropriateness.

The appropriateness is represented on an ordinal scale that uses integers from 1 to 9 grouped into three categories: 1, 2, or 3 are in the category "usually not appropriate" where the harms of doing the procedure outweigh the benefits; and 7, 8, or 9 are in the category "usually appropriate" where the benefits of doing a procedure outweigh the harms or risks. The middle category, designated "may be appropriate", is represented by 4, 5, or 6 on the scale. The middle category is when the risks and benefits are equivocal or unclear, the dispersion of the individual ratings from the group median rating is too large (i.e., disagreement), the evidence is contradictory or unclear, or there are special circumstances or subpopulations which could influence the risks or benefits that are embedded in the variant.

The ratings assigned by each panel member are presented in a table displaying the frequency distribution of the ratings without identifying which members provided any particular rating. To determine the panel's recommendation, the rating category that contains the median group rating without disagreement is selected. This may be determined after either the first or second rating round. If there is disagreement after the second rating round, the recommendation is "May be appropriate."

This modified Delphi method enables each panelist to articulate his or her individual interpretations of the evidence or expert opinion without excessive influence from fellow panelists in a simple, standardized, and economical process.

1c.23 Grade Assigned to the Recommendation: No grade for AUA best practice statement, NCCN grade 2A
AUA 2013 Guideline:
No grade or definition has been provided for this guideline.

NCCN 2016 Guideline:

Level of evidence assigned: Category 2A

ACR 2012 Appropriateness Criteria:
Appropriateness Rating Assigned: 8

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: Moderate **1c.26** Quality: Moderate**1c.27** Consistency**:** Moderate

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0389_Evidence_MSF5.0_Data-635278494960508026-635932939997439723.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Multiple studies have indicated that a bone scan is not clinically necessary for staging prostate cancer in men with a low risk of recurrence and receiving primary therapy. For patients who are categorized as low-risk, bone scans are unlikely to identify their disease. Furthermore, bone scans are not necessary for low-risk patients who have no history or if the clinical examination suggests no bony involvement. Less than 1% of low-risk patients are at risk of metastatic disease.

While clinical practice guidelines do not recommend bone scans in low-risk prostate cancer patients, overuse is still common. An analysis of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it (1). The analysis also found that the use of bone scans in low-risk patients leads to an annual cost of $4 million dollars to Medicare. The overuse of bone scan imaging for low-risk prostate cancer patients is a concept included on the American Urological Association's list in the Choosing Wisely Initiative as a means to promote adherence to evidence-based imaging practices and to reduce health care dollars wasted (2). This measure is intended to promote adherence to evidence-based imaging practices, lessen the financial burden of unnecessary imaging, and ultimately to improve the quality of care for prostate cancer patients in the United States.

Citations:

1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

2. American Urological Association. A routine bone scan is unnecessary in men with low-risk prostate cancer. Choosing Wisely Initiative. Released February 21, 2013. Accessed February 25, 2016.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included).*

This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI
Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 EHR Performance Rate:
Mean: 90.76%
Maximum: 100.00%
Minimum: 50.00%

01/01/2014 – 12/31/2014 Registry Performance Rate:
Mean: 90.24%
Minimum: 0.00%
Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients the last several years are as follows:
2010: 71.60%
2011: 90.50%
2012: 92.50%
2013: 88.50%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 8.2% of eligible professionals reported on Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients for claims, registry, and electronic health records. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

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.
An analysis conducted by Falchook and colleagues of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it. Given that low-risk patients have a less than 1% chance of developing a metastatic disease, the authors suggest that bone scan imaging in low-risk prostate cancer patients contributes to poor quality care and is a large contributor of health care dollars wasted in the United States (1). The literature recommends clinician education on guideline recommendations to spur improvement. These findings support the need for an NQF endorsed performance measure along with targeted initiatives to improve adherence to appropriate imaging.

Citation:
1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
While this measure is included in federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
African American/Black men have the highest incidence rate of prostate cancer in the United States and are more than twice as likely as white men to die from the disease (1). Between 2008-2012, the average annual prostate cancer incidence rate among African American men was 208.7 cases per 100,000 men, which was 70% higher than the rate in white men. Although prostate

cancer incidence and mortality rates have been declining in African American and white men since 1991, the incidence, prevalence, and death rates remain comparably higher among African American men as compared to white men (2).

An analysis of the SEER Medicare database conducted by Falchook and colleagues found that imaging overuse was associated with nonwhite race, higher comorbidity, and regional education and income measures (3). An additional analysis SEER Medicare database found there was regional variation in the use of bone scans in low- and intermediate- risk patients with the highest use in the Northeast and lowest use in the West (4).

Citations:

1. National Cancer Institute. Cancer health disparities. http://www.cancer.gov/about-nci/organization/crchd/cancer-health-disparities-fact-sheet. Accessed February 12, 2016.

2. American Cancer Society Cancer Facts and Figures for African Americans 2016-2018. http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-047403.pdf. Accessed February 26, 2016.

3. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

4. Falchook AD, Salloum RG, Hendrix LH, Chen RC. Use of bone scan during initial prostate cancer workup, downstream procedures, and associated Medicare costs. Int J Radiat Oncol Biol Phys. 2014;89(2):243-248.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, High resource use

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

220,800 new cases of prostate cancer were diagnosed in 2015 (1). The number of new cases of prostate cancer was 137.9 per 100,000 men per year, based on age-adjusted cases and deaths between 2008-2012. In 2012, there were an estimated 2,795,592 men living with prostate cancer in the United State. Prostate cancer is the third most common type of cancer in the United State and accounts for 13.3% of all new cancer cases. Approximately 14% of men will be diagnosed with prostate cancer at some point in their lifetime, based on 2010-2012 data (2).

The annual estimated cost to Medicare of all bone scans for prostate cancer patients is $19,300,000 including $9,300,000 for low and intermediate-risk patients. An additional $2,000,000 is spent annually on downstream imaging bone scans for low- and-intermediate risk patients (3). The annual Medicare of bone scans for low-risk prostate cancer patients alone is $4,000,000 (4).

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Siegel RL, Miller KD, Jemal A. Cancer statistics 2015. Ca Cancer J Clin. 2015;65:5-29.

2. SEER Cancer Statistics Factsheets: Prostate Cancer. National Cancer Institute. Bethesda, MD,http://seer.cancer.gov/statfacts/html/prost.html

3. Falchook AD, Salloum RG, Hendrix LH, Chen RC. Use of bone scan during initial prostate cancer workup, downstream procedures, and associated Medicare costs. Int J Radiat Oncol Biol Phys. 2014;89(2):243-248.

4. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not applicable. Not a PRO-PM.

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 subcriteria 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):*
Cancer, Cancer : Prostate

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse

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.)*
Measure specifications are included as an attachment with this submission. Additional measure details may be found at:http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/eCQM_Library.html Value sets at https://vsac.nlm.nih.gov

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:** EP_eCQM_ValueSets_CMS129v6_NQF0389_02182016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The measure description, denominator statement, denominator details and value sets, were revised based on updated risk strata to be consistent with NCCN guidelines.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who did not have a bone scan performed at any time since diagnosis of prostate cancer

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Once for each procedure for treatment of prostate cancer (ie, interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy)during the 12-month reporting period

S.6. **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, 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.
For Registry:
To submit the numerator option for patients who did not have a bone scan performed at any time since diagnosis of prostate cancer, report the following CPT Category II code:

3270F – Bone scan not performed prior to initiation of treatment nor at any time since diagnosis of prostate cancer

For EHR Specifications:
HQMF eMeasure developed and is included in this submission.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence, receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Definitions:
Risk Strata Definitions: Very Low, Low, Intermediate, High, or Very High-
Very Low Risk - PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1c; AND presence of disease in fewer than 3 biopsy cores; AND <= 50% prostate cancer involvement in any core; AND PSA density <= 0.15 ng/mL/cm3.
Low Risk - PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1 to T2a.
Intermediate Risk - PSA 10 to 20 ng/mL; OR Gleason score 7; OR clinical stage T2b to T2c. Note: Patients with multiple adverse factors may be shifted into the high risk category.
High Risk - PSA > 20 ng/mL; OR Gleason score 8 to 10; OR clinically localized stage T3a. Note: Patients with multiple adverse factors may be shifted into the very high risk category.
Very High Risk - Clinical stage T3b to T4; OR primary Gleason pattern 5; OR more than 4 cores with Gleason score 8 to 10. (NCCN, 2016)

External beam radiotherapy - external beam radiotherapy refers to 3D conformal radiation therapy (3D-CRT), intensity modulated radiation therapy (IMRT), stereotactic body radiotherapy (SBRT), and proton beam therapy.

Note: Only patients with prostate cancer with low risk of recurrence will be counted in the denominator of this measure

For Registry:
Any male patient, regardless of age
AND
Diagnosis for prostate cancer (ICD-9-CM): 185
Diagnosis for prostate cancer (ICD-10-CM): C61
AND
Patient encounter during the reporting period (CPT): 55810, 55812, 55815, 55840, 55842, 55845, 55866, 55873, 55875, 77427, 77435, 77772, 77778, 77799
AND
Report the following CPT Category II Code to identify the risk of recurrence:
3271F: Low risk of recurrence, prostate cancer

For EHR:
HQMF eMeasure developed and is included in this submission.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons)

Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician)

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients, exceptions may include medical reason(s) (eg, documented pain, salvage therapy, other medical reasons) or system reason(s) (eg, bone scan ordered by someone other than reporting physician). Where examples of exceptions are included in the measure language, value sets for these examples are developed and included in the eMeasure. Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Additional details by data source are as follows:

For Registry:
Append modifier to CPT Category II code:
3269F with 1P - Documentation of medical reason(s) for performing a bone scan (including documented pain, salvage therapy, other medical reasons)

Append modifier to CPT Category II code:
3269F with 3P - Documentation of system reason(s) for performing a bone scan (including bone scan ordered by someone other than reporting physician)

For EHR:
HQMF eMeasure developed and is included in this submission.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer and have included these variables as recommended data elements to be collected.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment or risk stratification

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).
2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.
3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: medical reason(s) (eg, documented pain, salvage therapy, other medical reasons) or system reason(s) (eg, bone scan ordered by someone other than reporting physician)]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable. The measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable. The measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Patient eligibility is determined by a set of defined criteria relevant to a particular measure. If data required to determine patient eligibility are missing, those patients/cases would be ineligible for inclusion in the denominator and therefore the patient/case would be deleted.

If data required to determine if a denominator eligible patient qualifies for the numerator (or has a valid exclusion/exception) are missing, this case would represent a quality failure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not applicable.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Clinician : Team

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinician Office/Clinic, Other
If other: Radiation Oncology Clinic/Department

S.28. **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. The measure is not a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0389_Avoidance_of_Overuse_of_Bone_Scans_Updated_Testing.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0389 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
Five practice sites representing various types, locations and sizes were identified to participate in testing the 3 PCPI/ASTRO/AUA-developed prostate cancer performance measures.
o Site A: hospital, multi-practice sites in urban, rural and suburban settings; 21 physicians; average 9600 oncology/prostate cancer patient visits per month for MD/NP assessment, chemo; submitted PQRS claims for one measure and utilized a full-fledged EHR.
o Site B: physician owned private practice, suburban setting; 4 physicians; average 48 oncology/prostate cancer patients seen

per day; submitted PQRS claims for one measure and utilized paper medical records.
o Site C: physician owned private practice, urban setting; 41 physicians; average 2500 oncology/prostate cancer patients seen per month; submitted PQRS claims for two measures and utilized a full-fledged EHR.
o Site D: academic, suburban setting; 9 physicians; average 240 oncology/prostate cancer patients seen per month; submitted PQRS claims for one measure and utilized paper and EHR.
o Site E: academic, urban setting; 14 physicians; average 250 oncology/prostate cancer patients seen per month; collected PQRS data on 3 measures and utilized a full-fledged EHR.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

Signal to Noise Ratio analysis data
The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

The data are for the time period January 2014 through December 2014 and cover the entire United States.

The total number of physicians reporting on this measure, via the registry reporting option, in 2014, is 131. Of those, 24 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 1,296 quality events. For this measure, 18.3 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 46 for the remaining 1,113 events. The range of quality reporting events for 24 physicians included is from 197 to 10. The average number of quality reporting events for the remaining 81.7 percent of physicians that aren't included is 1.2.

There were 1,113 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure and remained after exceptions were removed.

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)*:
PCPI Testing Project
Data abstracted from patient records were used to calculate inter-rater reliability for the measure.
94 patient records were reviewed.

Data analysis included:
- Percent agreement; and
- Kappa statistic to adjust for chance agreement.

Signal to Noise Ratio analysis data
Reliability of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.
Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

2a2.3 **Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
PCPI Testing Project

N, % Agreement, Kappa (95% Confidence Interval)
Overall Reliability: 94, 100%, Kappa is noncalculable*
Denominator Reliability: 94, 100%, Kappa is noncalculable*
Numerator Reliability: 94, 100%, Kappa is noncalculable*
Exceptions Reliability: 94, 100%, Kappa is noncalculable*

This measure demonstrates perfect reliability, as shown in results from the above analysis.

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

Signal to Noise Ratio analysis results
This measure has 0.84 reliability when evaluated at the minimum level of quality reporting events and 0.96 reliability at the average number of quality events.

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is very high.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
The evidence directly supports the specified measure. The measure specifically identifies the risk strata for whom bone scans are inappropriate. The guideline and best practice statement do not recommend bone scans for patients included in the low risk strata.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.

The panel members are as follows:

Ian Thompson, MD (Co-Chair, urology)
Steven Clauser, PhD (Co-Chair, methodology)
Peter Albertsen, MD (urology)
Colleen Lawton, MD (radiation oncology)
Charles Bennett, MD, PhD, MPP (clinical oncology)
W. Robert Lee, MD, MS, Med (radiation oncology)
Michael Cookson, MD (urology)
Peter A. S. Johnstone, MD, FACR (radiation oncology)
Gregory W. Cotter, MD (radiation oncology)
David F. Penson, MD, MPH (urology)
Theodore L. DeWeese, MD (radiation oncology)
Stephen Permut, MD (family medicine)
Mario Gonzalez, MD (pathology)
Howard Sandler, MD (radiation oncology)
Louis Kavoussi, MD (urology)
Bill Steirman, MA (consumer representative)
Eric A. Klein, MD (urology)
John T. Wei, MD (urology)
Carol Wilhoit, MD (health plan representative)

2016 Face validity assessment
The expert panel included 17 members. Panel members were comprised of experts from the PCPI Measures Advisory Committee. The

list of expert panel members is as follows:
Joseph Drozda, MD, FACC (Chair)
Richard Bankowitz, MD, MBA, FACP
Heidi Bossley, MSN, MBA
John Easa, MD, FIPP
Christine Goertz, DC, PhD
Jeff Jacobs MD, FACS, FACC, FCCP
Yosef Khan MD, MPH, PhD, MACE
Dianne Jewell, PT, DPT, PhD, FAACVPR
Scott T. MacDonald, MD
Mark Metersky, MD
Michael O'Dell, MD, MS, MSHA, FAAFP
Martha Radford, MD, FACC, FAHA
Amy Sanders, MD, MS
David Seidenwurm, MD
Shannon Sims, MD, PhD
Jessie Sullivan, MD
Karen Johnson (NQF Liaison) - RECUSED

2b2.2 **Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert Work Group and the measures adjusted as needed. Other external review groups (i.e. focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity has been quantitatively assessed for this measure. Specifically, the Prostate Cancer Work Group members were asked to empirically assess face validity of the measure. The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.

Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1=Disagree; 3=Neither Disagree nor Agree; 5=Agree

2016 Face validity assessment
Face validity of the measure score as an indicator of quality was systematically assessed as follows.
After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:
The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.
Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

---

To satisfy NQF's ICD-10 Conversion Requirements, we are providing the information below:

- NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM
  Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.
- NQF ICD-10-CM Requirement 2: Coding Table
  See attachment in S.2b
- NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes

The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the

ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
The results of the expert panel rating of the validity statement were as follows: N = 13; Mean rating = 4.6

Percentage in the top two categories (4 and 5): 92.31%

Frequency Distribution of Ratings

1 – 0
2 – 1
3 – 0
4 – 2
5 – 10

2016 Face validity assessment
The results of the expert panel rating of the validity statement were as follows: N = 10; Mean rating = 3.8 and 80% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.
Frequency Distribution of Ratings
1 – 0 responses (Strongly Disagree)
2 – 2 responses
3 – 0 responses (Neither Agree nor Disagree)
4 – 6 responses
5 – 2 responses (Strongly Agree)

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
- 94 patient records were reviewed for this measure.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

2014 PQRS Registry Data
The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

The data are for the time period January 2014 through December 2014 and cover the entire United States.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Exceptions were analyzed for frequency and variability across providers.

Exceptions are necessary to account for those situations when it is medically appropriate for a patient to have a bone scan. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for several reasons. Rather than specifying an exhaustive list of explicit reasons for exception for this measure, the measure developer relies on clinicians to link the exception with a specific reason for the decision to order a bone scan required for a patient.

Some have indicated concerns with exception reporting including the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234.

Exceptions were analyzed for frequency across providers.

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
PCPI Testing Project
N, % Agreement, Kappa (95% Confidence Interval)
Exceptions Reliability: 94, 100%, Kappa is noncalculable*

This measure demonstrates almost reliability, as shown in results from the above analysis.

The exception rate for this measure is 6.4%

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2014 PQRS Registry Data
Amongst the 24 physicians with the minimum (10) number of quality reporting events, there were a total of 183 exceptions reported. The average number of exceptions per physician in this sample is 7.6. The overall exception rate is 14.1%.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
This measure is not risk adjusted.

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
This measure is not risk adjusted.

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

Not applicable

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
As a process measure, no risk adjustment is necessary.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
- 94 patient records were reviewed for this measure.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

CMS Physician Quality Reporting Initiative:
Clinical Condition and Measure: #102
14,484 patients were reported on for the 2008 program, the most recent year for which data are available

In 2009 the following was reported for this measure:
Eligible Professionals: 8,138
Professionals Reporting >=1 Valid QDC: 471
% Professionals Reporting >=1 Valid QDC: 5.79%
# Professionals Satisfactorily Reporting: 163
% Professionals Satisfactorily Reporting: 34.61%

2014 PQRS Registry Data
The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

The data are for the time period January 2014 through December 2014 and cover the entire United States.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
PCPI Testing Project
Data analysis performed on the measure included:
Average measure performance rate overall and by site, performance rate range by site and overall standard deviation for the measure.

CMS Physician Quality Reporting Initiative:
The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

2014 PQRS Registry data
Measures of central tendency, variability, and dispersion were calculated.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PCPI Testing Project
Measure rate without exceptions: N= 94 Mean = 47.9% Standard Deviation= 0.5022
The performance rate by site is as follows, where n is the number of performance events by site:

A	0.1670	n=30
B	0.5710	n=7
C	0.5000	n=30
D	0.7780	n=27

The performance rate range is .6110. Although this study captured performance on 94 events, the data were not captured at the physician level, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a

meaningful calculation of variation in performance across organizations due to the small sample size of sites (n=4) in this study.

CMS Physician Quality Reporting Initiative
This measure was used in the 2008-2011 CMS Physician Quality Reporting Initiative Claims and Registry options and group reporting option available in 2011.

There is a gap in care as shown by this 2008 data, the only year for which distribution by quartile/decile is available.

84.31% of patients reported on did not meet the measure.

10th percentile: 0.00%
25th percentile: 0.00%
50th percentile: 1.36%
75th percentile: 35.00%
90th percentile: 77.63%

The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 35.00 and indicates that 50% of physicians have performance on this measure ranging from 0.00% and 35.00%. A quarter of reporting physicians have performance on this measure greater than 35.00%, while a quarter have performance on this measure at 0.00%.

2014 PQRS Registry data
Based on the sample of 24 included physicians, the mean performance rate is 0.89 the median performance rate is 1.0 and the mode is 1.0. The standard deviation is 0.20. The range of the performance rate is 0.49, with a minimum rate of 0.51 and a maximum rate of 1.00. The interquartile range is 0.12 (0.88 – 1.00).

The range of performance from 0.51 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
PCPI Testing Project
- 41 Medicare patient records of the 94 patient records were reviewed.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
PCPI Testing Project
Parallel forms reliability testing was performed. PQRS claims were reviewed and compared to a manual review of claims information.

Data analysis included:
- Percent agreement

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
PCPI Testing Project
N, % Agreement
41, 100%

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one's ancestry) and language need (a rating of spoken English language proficiency of less than very well and one's preferred language for health-related encounters)."(2)

References:
(1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.

(2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

2.1-2.3 **Supplemental Testing Methodology Information:**

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

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

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

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.**
Attachment **Attachment:** NQF_389_Feasibility_Assessment_and_Bonnie_attachment.pdf

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.
The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA). Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting PQRS http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/pqrs/index.html Payment Program Meaningful Use Stage II https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/

4a.1. **For each CURRENT use, checked above, provide:**

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

1.Physician Quality Reporting System (PQRS) – Sponsored by the Centers for Medicare and Medicaid Services (CMS)

Purpose: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013. Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented a phased approach to public reporting performance information on the Physician Compare Web site. CMS announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in late 2017.

2.Meaningful Use Stage 2 (EHR Incentive Program) – Sponsored by the Centers for Medicare and Medicaid Services (CMS)

Purpose: The Medicare and Medicaid EHR Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology.

Eligibility for incentive payments for the "meaningful use" of certified EHR technology is established if all program requirements are met, including successful implementation and reporting of program measures, which include this measure, to demonstrate meaningful use of EHR technology.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI
Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 EHR Performance Rate:
Mean: 90.76%
Maximum: 100.00%
Minimum: 50.00%

01/01/2014 – 12/31/2014 Registry Performance Rate:
Mean: 90.24%
Minimum: 0.00%
Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients the last several years are as follows:
2010: 71.60%
2011: 90.50%
2012: 92.50%
2013: 88.50%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 8.2% of eligible professionals reported on Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients for claims, registry, and electronic health records. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

4b.2. **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.**
While the PCPI creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1.Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences at this time, but we take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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)**

0390 : Prostate Cancer: Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients
1853 : Radical Prostatectomy Pathology Reporting

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

5a. Harmonization

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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The related measure 1853, Radical Prostatectomy Pathology Reporting, addresses the percentage of radical prostatectomy pathology reports that include the pT category, the pN category, the Gleason score and a statement about margin status, which is a different action than measure 0389. The two measures do not share similar target populations and address different aspects of prostate cancer care. The related measure 0390, Prostate Cancer: Adjuvant Hormonal Therapy for High Risk or Very High Risk Prostate Cancer Patients addresses the use of adjuvant hormonal therapy and external beam radiation therapy in high-risk prostate cancer patients which is a different quality action from measure 0389. The two measures do not share similar target populations and address different aspects of prostate cancer care.

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.
No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** PCPI
Co.2 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-
Co.3 **Measure Developer if different from Measure Steward:** PCPI
Co.4 **Point of Contact:** Diedra, Gray, diedra.gray@ama-assn.org, 312-464-4904-

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.
Ian Thompson, MD (Co-Chair, urology)

Steven Clauser, PhD (Co-Chair, methodology)
Peter Albertsen, MD (urology)
Colleen Lawton, MD (radiation oncology)
Charles Bennett, MD, PhD, MPP (clinical oncology)
W. Robert Lee, MD, MS, Med (radiation oncology)
Michael Cookson, MD (urology)
Peter A. S. Johnstone, MD, FACR (radiation oncology)
Gregory W. Cotter, MD (radiation oncology)
David F. Penson, MD, MPH (urology)
Theodore L. DeWeese, MD (radiation oncology)
Stephen Permut, MD (family medicine)
Mario Gonzalez, MD (pathology)
Howard Sandler, MD (radiation oncology)
Louis Kavoussi, MD (urology)
Bill Steirman, MA (consumer representative)
Eric A. Klein, MD (urology)
John T. Wei, MD (urology)
Carol Wilhoit, MD (health plan representative)

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2007

Ad.3 **Month and Year of most recent revision:** 09, 2015

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, 2017

Standards Development Organisation (IHTSDO). ICD-10 is copyright 2015 World Health Organization. All Rights Reserved. Ad.7 **Disclaimers:** Please see the copyright statement above in AD.6 for disclaimer information.
Ad.8 **Additional Information/Comments:**

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 #: 0390
Measure Title: Prostate Cancer: Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients
Measure Steward: American Urological Association
Brief Description of Measure: Percentage of patients, regardless of age, with a diagnosis of prostate cancer at high or very high risk of recurrence receiving external beam radiotherapy to the prostate who were prescribed adjuvant hormonal therapy (GnRH [gonadotropin-releasing hormone] agonist or antagonist)
Developer Rationale: The use of adjuvant hormonal therapy following external beam radiotherapy is a well-established standard of care for high-risk prostate cancer patients. Multiple large studies have shown that men who receive adjuvant hormonal therapy following external beam radiotherapy can live longer and have a lower risk of recurrence than men who receive radiotherapy alone. In addition, a cost-analysis conducted found that the use of adjuvant hormonal therapy and external beam radiotherapy is cost-effective and adds quality-adjusted life years for patients (1).

Data from several sources indicates that while utilization rates of adjuvant hormonal therapy and external beam radiotherapy have increased, they still remain suboptimal. One study analyzing the CaPSURE database, a provider-based registry, found that the utilization of adjuvant hormonal therapy and external beam radiotherapy for high-risk patients has increased to 80% throughout the past two decades, yet utilization rates have plateaued since 2000 (2). There is rising concern about undertreatment of high-risk prostate cancer patients (3). This suggests greater outreach and education are needed to improve outcomes in care.

Citation:
1. Satish K, Shelly M, Harrison C, et al. Neo-adjuvant and adjuvant hormone therapy for localised and locally advanced prostate cancer. Cochrane Database Syst Rev. 2006; (4): CD006019. DOI: 10.1002/14651858.CD006019.pub2. Accessed at: http://www.cochrane.org/CD006019/PROSTATE_neo-adjuvant-and-adjuvant-hormone-therapy-for-localised-and-locally-advanced-prostate-cancer

2. Cooperberg MR, Janet Cowan, J, Broering JM, et al. High-Risk Prostate Cancer in the United States, 1990-2007. World J Urol. 2008; 26(3): 211–218. doi:10.1007/s00345-008-0250-7.

3. Cooperberg MR, Broering JM, Caroll, PR. Time trends and local variation in primary treatment of localized prostate cancer. J Clin Oncol. 2010;28(7):1117-1123.

Numerator Statement: Patients who were prescribed adjuvant hormonal therapy (GnRH [gonadotropin-releasing hormone] agonist or antagonist)
Denominator Statement: All patients, regardless of age, with a diagnosis of prostate cancer at high or very high risk of recurrence receiving external beam radiotherapy to the prostate
Denominator Exclusions: AUA methodology uses three categories of reasons for which a patient may be excluded from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For this measure, exceptions for not prescribing/administering adjuvant

hormonal therapy may include medical reason(s) (eg, salvage therapy) or patient reason(s). Although this methodology does not require the external reporting of more detailed exception data, the AUA recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The AUA also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement. For example, it is possible for implementers to calculate the percentage of patients that physicians have identified as meeting the criteria for exception. Additional details by data source are as follows:

Documentation of medical reason(s) for not prescribing/administering adjuvant hormonal therapy (eg, salvage therapy)

Documentation of patient reason(s) for not prescribing/administering adjuvant hormonal therapy

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Aug 09, 2012

Maintenance of Endorsement-- Preliminary Analysis

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 *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.

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:

- The evidence for this measure was based an American Urological Association (AUA) Standard from the Guideline for the management of clinically localized prostate cancer: 2007 update and a clinical practice guideline from the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011:
  - AUA Standard: When counseling patients regarding treatment options, physicians should consider the following: Based on results of two randomized controlled clinical trials, the use of adjuvant and concurrent hormonal therapy may prolong survival in the patient who has opted for radiotherapy. High-risk patients who are considering specific treatment options should be informed of findings of recent high-quality clinical trials, including that: For those considering external beam radiotherapy, use of hormonal therapy combined with conventional radiotherapy may prolong survival. **Level of evidence: Standard**. [*AUA Guideline Statement Definitions: Standard: A guideline statement is a standard if: (1) the health outcomes of the alternative interventions are sufficiently well known to permit meaningful decisions, and (2) there is virtual unanimity about which intervention is preferred*.]
  - NCCN guideline recommendation: There are several treatment options for patients with high-risk disease. The preferred treatment is 3D-CRT/IMRT with daily IGRT in conjunction with long-term ADT;

ADT alone is insufficient. In particular, patients with low volume, high grade tumor warrant aggressive local radiation combined with typically 2-3 years of ADT. **Level of evidence: NCCN Category 1** [*The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.]*

- In 2012, the Committee state that the evidence provided is high level and supportive of the measure focus.

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 updates to the guidelines, and included a Cochrane Review to the body of evidence.

- The developer provided the NCCN 2016 Guideline update:
  - Men with prostate cancer that is clinical stage T3a, Gleason score 8 to 10, or PSA level greater than 20 ng/mL are categorized by the panel as high risk. Patients with multiple adverse factors may be shifted to the very high-risk category. [See detailed risk strata below]. The preferred treatment is EBRT [external beam radiation therapy] in conjunction with 2 to 3 years of neoadjuvant/concurrent/adjuvant ADT [androgen deprivation therapy] (category 1); ADT alone is insufficient. In particular, patients with low-volume, high-grade tumor warrant aggressive local radiation combined with typically 2 or 3 years of neoadjuvant/concurrent/adjuvant ADT. Fit men in the high-risk group can consider 6 cycles of docetaxel without prednisone after EBRT is completed and while continuing ADT. The combination of EBRT and brachytherapy with our without neoadjuvant/concurrent/adjuvant ADT, is another primary treatment option. However, the optimal duration of ADT in this setting remains unclear.
  - Patients at very high risk (locally advanced) are defined by the NCCN Guidelines as men with clinical stages T3b to T4, primary Gleason pattern 5, or more than 4 biopsy cores with Gleason score 8 to 10. The options for this group include: 1) EBRT and long-term ADT (category 1); 2) EBRT plus brachytherapy with or without long-term ADT; 3) EBRT plus ADT and docetaxel; 4) radical prostatectomy plus PLND in selected patients with no fixation to adjunct organs; or 5) ADT for patients not eligible for definitive therapy.
    - The developer also included the risk strata for these recommendations.
  - The **level of evidence** for these recommendations is **NCCN's Category 1** [*Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate*].

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as high-level evidence (Box 6) → Rate as Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *Is the Committee willing to accept the prior evaluation? The updated evidence supports the measure focus and has a stronger level of 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.

- The developer provided the following PQRS Registry and Part B Claims data from January 1, 2014 – December 31, 2014:

	Registry Performance Rate
Mean	93.82%
Minimum	16.67%
Maximum	100.00%

PQRS Experience Report

	Average Performance Rates
2010	79.60%
2011	93.50%
2012	91.10%
2013	95.40%

- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).
- The developer provided additional data from the literature.

Disparities

- The developer stated that federal reporting programs have not yet made disparities data available to analyze and report. Disparities data from the measure as specified is encouraged for endorsement maintenance; this measure has been endorsed since 2008.
- The developer presents evidence from the literature that describes racial/ethnic differences of incidence rate of prostate cancer (between 2008 – 2012 the African American men was 208.7 cases per 100,000 men, which was 70% higher than the rate in white men) and differences in treatement (African American men are more likely to receive non-surgical treatment than white men. White men were 25% less likely than African American men to receive radiation therapy and are 48% less likely to receive hormonal therapy).

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement ?*
- *Does the data presented demonstrate 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**

Rationale: Insufficient information provided to determine if a performance gap exists.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The recommendation for the addition of ADT to XRT in high risk prostate cancer is based on the results of two RCT and was given the highest level of evidence score by both AUA (standard) and NCCN (Category 1) guidelines. Since last review, 2016 updated NCCN guidelines maintain a category 1 rating for the recommendation. Risk strata definitions have been updated to be consistent with this guideline. Further, an interval Cochrane Review is cited in support of high level evidence for this measure. Rating: MODERATE

****Measure continues to be included in guidelines from the NCCN 2016 with level 1 evidence, a cochrane review and the AUA. This is well supported by evidence and rationale.****

1b. Performance Gap

Comments:

**Performance in 2014 ranged from 16 to 100% with mean of 94% based on PQRS and claims data. This has increased from 2010 average rate of 79%, and remained stable since 2011.

Disparity data were not provided. literature citing increased morbidity and mortality for AA men was cited, showing that AA were 25% more likely to receive radiation, but 48% less likely to receive ADT (concerning).**
The performance gap is the main area of uncertainty. The PQRS reporting demonstrated performance of 95.4% and 93.8%. The literature is referenced regarding disparities in treatment of african americans but a specific analysis was not included. Question whether there is opportunity to further close the performance gap.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): electronic clinical data, registry data. This is not an e-Measure.

Specifications:

- The level of analysis is at the clinician-level and is tested at the Ambulatory Care and or Radiation Oncology Clinic/Department settings.
- The numerator is 'Patients who were prescribed adjuvant hormonal therapy (GnRH [gonadotropin-releasing hormone] agonist or antagonist)'.
    - The developer defines prescribed as "Includes patients who are currently receiving medication(s) that follow the treatment plan recommended at an encounter during the reporting period, even if the prescription for that medication was ordered prior to the encounter".
- The denominator includes 'All patients, regardless of age, with a diagnosis of prostate cancer at high or very high risk of recurrence receiving external beam radiotherapy to the prostate'.
    - The developer defines 'High' and 'High Risk' as:
        - High Risk: PSA > 20 ng/mL; OR Gleason score 8 to 10; OR clinically localized stage T3a. Note: Patients with multiple adverse factors may be shifted into the very high risk category.
        - Very High Risk: Clinical stage T3b to T4; OR primary Gleason pattern 5; OR more than 4 cores with Gleason score 8 to 10. (NCCN, 2016)
- Denominator exclusions include:
    - Documentation of medical reason(s) for not prescribing/administering adjuvant hormonal therapy (eg, salvage therapy)
    - Documentation of patient reason(s) for not prescribing/administering adjuvant hormonal therapy
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The measure is not risk-adjusted.
- A calculation algorithm is provided and describes the process of calculating the measure.
- Missing data either delete a case from the denominator or represent a quality failure.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability was conducted on 91 patient records from 2010. Chart and data auditing occurred in

2011. The developer provided the percent agreement and kappa statistic (95% CI):
- Overall Reliability: N=91, 98.9%, 0.972 (0.916-1.000)
- Denominator Reliability: N=100%, Kappa is noncalculable*
- Numerator Reliability: N=100%, 0.971 (0.913-1.000)
- Exceptions Reliability: N=100%, Kappa is noncalculable*
- *Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

Describe any updates to testing

- Reliability of the measure score was not presented in prior submission(s), reliability testing of the measure score has been conducted this review.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 2014 Registry data from PQRS. A total of 86 physicians reported on this measure in 2014. Of those, 20 physicians had 515 patient records with all the required data elements and a **minimum of 10 quality reporting events**. The **average number of quality reporting events** (after exceptions were removed) was **21.5**.
- There were 430 patients included in this reliability testing and analysis of this measure.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability.

Results of reliability testing:

- Reliability at the at the minimum level of quality reporting events (10) was **0.73** and **0.85** at the average number of quality events (21.5).

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Appropriate method used/small sample size(Box 5) →Moderate reliability statistic and scope (Box 6) →Moderate

Questions for the Committee:

- *Is the test sample of 20 physicians adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing

- Additional empirical validity testing of the measure score has been conducted since the last review of this measure.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

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

Validity testing method

- The developer conducted new face validity testing with input from an expert panel including 21 members. The panel was comprised of experts from the AUA Committee on Quality Improvement and Patient Safety.

Validity testing results:

- **100%** (15) of the respondents either agreed or strongly agreed that this measure can accurately distinguish good and poor quality.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer notes several exclusions, as follows:
  - Documentation of medical reason(s) for not prescribing adjuvant hormonal therapy (eg, salvage therapy)
  - Documentation of patient reason(s) for not prescribing adjuvant hormonal therapy
- The developer reported that there were a total of 204 exceptions reported amongst the 20 physicians with the minimum (10) number of quality reporting events. The average number of exceptions per physician was 10.2 and overall exception rate was 32.2%.
- Without the exclusions "the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives".
- The developer also states that they recommend physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. AUA also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Questions for the Committee:

- *Are the results from the exclusions analysis a threat to validity? Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the seemingly high number of exclusions reasonable?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer calculated measures of central tendency, variability, and dispersion. Based on a sample of 20 physicians:
  - Mean performance rate is 0.93
  - Median performance rate is 1.00, and the mode is 1.00 - *this data provided by the developer may an error*
  - Standard deviation is 0.15
  - Range of the performance rate is 0.57, with a minimum rate of 0.43 and a maximum rate of 1.00
  - Interquartile range is 0.06 (0.94 – 1.00)

Question for the Committee:

- *Does a sample size of 20 physicians demonstrate statistically significant and meaningful differences in quality across physicians?*

2b6. Comparability of data sources/methods:

- Measure is not specified for more than one data source; comparability of data sources is not needed.

2b7. Missing Data

- The developer addressed how to handle missing data in S22; If data required to determine if a denominator eligible patient qualifies for the numerator (or has a valid exclusion/exception) are missing, this case would represent a quality failure.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Threats to validity mostly assessed (Box 2) →Empirical validity testing (Box 3)→ Face validity assessed (Box 4)→ Agreement measure can be used to distinguish quality (Box 5)→Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The denominator is clearly stated and reliably obtained from standard captured data: all HR and VHR receiving XRT The numerator is "patients who were prescribed adjuvant hormonal therapy" within 12 month reporting period. All appropriate codes are included and the measure can be consistently implemented.

The reliability of abstraction is high, except that the exceptions are not clearly defined, being at the discretion of the physician as related to clinical and patient reasons. This does not necessarily represent an issue with the measure as it allows flexibility as to the reason.

The specification are consistent with available evidence.

They are consistent with the evidence.

2a2. Reliability Testing

Comments:

Measure score testing used. 2014 PQRS dataset. 86 physicians reported with 20 meeting minimum 10 events. Of these 20, reliability was 0.73 at minimum (10) and 0.85 at average (21.5). Suggests that will support sufficient reliability.

Only 20 of 84 physicians had enough events and required elements for inclusion. Of these, a reliability of 0.73 was near the lower cut off of 0.7 for acceptability.

2b2. Validity Testing

Comments:

Face validity measured using expert panel of 21 (QUA quality committee). Of 15 respondents, 100% agreed or strongly agreed that the measure could distinguish good and poor quality care.

Of 15 committee respondents, 100% agreed or strongly agreed as to its face validity. It appears consistent.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Among the 20 physicians reviewed in PQRS, 203 exceptions were reported with average of 10 per physician and an overall exception rate of 32.2%. Documentation of specifics is recommended, though not reported. Systematic review and analysis is recommended. Without more granular data, difficult to assess nature of exclusion rate/appropriateness. unexplained exclusions are counted as failure. MODERATE

The exception rate is high at 32.2% but understandable in light of the patient population.

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.

- This measure is based on clinical registry data and all data elements are available in electronic sources.
- The Measure, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

Registry data easily and objectively abstracted.

This data would be expected to be available within the record, although not necessarily at the time of decision making. As long as the available information is taken into account, feasibility appears appropriate.

Criterion 4: Usability and Use

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

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details:

- Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS). The developer stated that CMS announced that there are plans to make all PQRS individual EP level PQRS measures available for public reporting on Physician Compare in late 2017.
- AUA Quality (AQUA) Registry - designed to measure and report healthcare quality and patient outcomes.
- This measure has been endorsed since 2008 - per NQF criteria, performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results:

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation

- The developer reports no additional difficulties or unexpected findings or benefits, apart from those included throughout the submission form.

Potential harms: The developer reports no unintended consequence were noted.

Feedback :

- In 2012, The Steering Committee found that this is a prevalent condition with a level of mortality that renders it a public health priority. The measure is supported by two randomized controlled trials, bolstered by expert opinion. The measure should be able to be reliably ascertained with EHR inputs.

***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

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

**Will be publicly reported beginning in 2017. Currently used in at least 1 accountability application.
No unintended consequences noted. Benefits outweigh. Continued improvement in use of adjuvant ADT in this population decrease progression and morbidity from prostate cancer and limit sub-optimal application or radiation therapy.**
The applicability as a performance measure is clear as is teh ability to drive improvements in care. Main question is the degree to which a gap continues to exist.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0220 : Adjuvant hormonal therapy

- 0389 : Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients
- 1853 : Radical Prostatectomy Pathology Reporting

Harmonization

- Not Harmpnozed with any of the measure listed.

Pre-meeting public and member comments

- As the measure steward, the American Urological Association supports this important measure.

NQF #: 0390 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)*:

The process of treating high risk prostate cancer patients with combination external beam radiotherapy and adjuvant hormone therapy is linked to improved outcomes, including prolonged survival.

1c.2-3 **Type of Evidence** *(Check all that apply)*:

Clinical Practice Guideline

Clinical Practice Guideline

Systematic Review

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The AUA, NCCN guidelines, and Cochrane Review recommend adjuvant hormonal therapy with radiotherapy for high risk prostate cancer patients, for prolonged survival. The measure captures patients receiving external beam radiotherapy in the denominator, and adjuvant hormonal therapy being prescribed in the numerator. Therefore, the evidence directly relates to the specified measure.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: The AUA guideline cites two randomized controlled clinical trials, within the recommendation regarding the use of adjuvant and concurrent hormonal therapy.

The description of the evidence review in the NCCN guideline did not address the overall quantity of studies in the body of evidence. However, 223 articles are cited in NCCN´s prostate cancer guideline´s reference section.

Cochrane Review:

Four randomized controlled trials are included in the review.

NCCN 2016 Guideline:

Information regarding the total number of studies and type of study designs included in the body of evidence is not

available. However, the guideline cites 1 observational study in support of the recommendation statement.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The AUA guideline cites two randomized controlled clinical trials, within the recommendation regarding the use of adjuvant and concurrent hormonal therapy.

The quality of the body of evidence supporting the NCCN guideline recommendation is summarized according to the NCCN categories of evidence and consensus as being based on "high level evidence (e.g. randomized controlled trials)."

Cochrane Review:

The review does not include an overall estimate of the quality of evidence across studies. However, the studies included are considered to be of sufficient quality to assess the use of adjuvant hormonal therapy and radiotherapy. The review also states that when considering the results of the review it is important to note inter-trial differences that relate to the heterogeneity of the results in some instances. Such variations include the duration of adjuvant treatment, the timing of radiotherapy, and different baseline population demographics across studies.

NCCN 2016 Guideline:

The guideline does not include an overall estimate of benefit from the body of evidence, however, the NCCN guideline recommendations are summarized according to NCCN categories of evidence and consensus as being based upon “high-level evidence.”

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The AUA guideline does not make any explicit statement regarding the overall consistency of the results across studies.

Although there is no explicit statement regarding the overall consistency of results across studies in the NCCN guideline, the recommendation received uniform NCCN consensus that the intervention is appropriate.

Cochrane Review:

The survival data from the 4 studies was pooled and analyzed. The review states: “At both 5 years (comparison 6: outcome 1) and 10 years (comparison 6: outcome 2), the test for overall effect was significantly in favour of adjuvant hormones (OR 1.29, 95% CI, 1.07 to 1.56, P=0.007 for 5 years and OR 1.44 95% CI 1.13 to 1.184 P=0.003 for 10 years). Heterogeneity was P=0.03 and P= 0.07, respectively.” The disease-specific survival data was pooled and analyzed for 2 of the 4 studies. The review states: “On pooling the data for the other two studies, the overall treatment effect was significantly in favour of adjuvant therapy (OR 2.10, 95% CI 1.53 to 2.88, P < 0.00001); however, significant heterogeneity was evident. (P=0.01) (comparison 6: outcome 3). The disease-free survival rates for 4 studies was pooled and analyzed. The review states: “The overall OR of .91 (95% CI 1.1.6 to 2.23) was significantly in favour of treatment (P<0.0001) though there was heterogeneity (P<0.0001). Pooling the 10 year survival data available for two studies (Zagars 1988; Pilepich 2005) with a total of 1059 patients (heterogeneity P=.69) gave an overall OR of 1.96 (95% CI 1.49 to 2.56) and was statistically significant (P <0.00001) in favour of the treatment arm (comparison 6: outcome 4).”

NCCN 2016 Guideline:

The guideline does not provide the consistency of results across studies, however, the recommendation received uniform NCCN consensus that the intervention is appropriate.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

The AUA guideline explains that the use of hormonal therapy may be associated with an increased risk of cardiovascular

disease and diabetes and that the use of hormone therapy in men who are at risk for or who are already diagnosed with heart disease and/or diabetes may have their health negatively impacted. However, per the AUA and NCCN guidelines, the use of adjuvant hormone therapy with external beam radiotherapy is the optimal course of therapy for high risk patients. This treatment combination may also prolong survival for high risk prostate cancer patients.

Cochrane Review:

The systematic review pooled data for the overall meta-analysis and states: "The addition of adjuvant hormone therapy with radiotherapy resulted in a significant improvement in overall survival, disease-specific survival, and disease-free survival at 10 years. However, due to heterogeneity, caution should be exercised with considering the pooled estimate of overall and disease-specific survival data."

Due to a lack of individual data in studies it was not possible to pool the data on adverse events for adjuvant hormonal therapy and radiotherapy. However, the adverse effects studied include hot flushes, diarrhea, liver function abnormalities, rash, nausea, and sexual inactivity during treatment. The review did not include how the harms affect the net benefits of treatment.

NCCN 2016 Guideline:

While no harms were mentioned in the guideline, it is expected that the harms of adjuvant hormonal therapy would include hot flashes, decreased libido, erectile dysfunction, gynecomastia, osteoporosis, weight loss, loss of muscle mass, fatigue, anemia, and changes in blood cholesterol and blood glucose measurements. The guideline does not include an overall estimate of benefit from the body of evidence, however, the use of adjuvant hormonal therapy with external beam radiotherapy is considered the optimal course of therapy for high risk patients, and in combination, may prolong survival for high-risk prostate cancer patients.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

Cochrane Review:

The Review did not provide a grade for an overall quality of the evidence.

NCCN 2016 Guideline:

Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: NCCN Prostate Cancer Panel

Andrew J. Armstrong, MD, ScM

Robert R. Bahnson, MD

Barry Boston, MD

J. Erik Busby, MD

Anthony Victor D´Amico, MD, PhD

James A. Eastham, MD

Charles A. Enke, MD

Thomas A. Farrington

Lauren Gallagher, RPh, PhD

Kristina M. Gregory, RN, MSN, OCN

Celestia S. Higano, MD, FACP

Maria Ho, PhD

Eric Mark Horwitz, MD

Philip W. Kantoff, MD

Mark H. Kawachi, MD

Michael Kuettel, MD, MBA, PhD

Richard J. Lee, MD, PhD

Gary R. MacVicar, MD

Arnold W. Malcolm, MD, FACR

Joan S. McClure, MS

David Miller, MD, MPH

James L. Mohler, MD

Elizabeth R. Plimack, MD, MS

Julio M. Pow-Sang, MD

Mack Roach, MD

Eric Rohren, MD, PhD

Stan Rosenfeld

Dorothy Shead, MS

Sandy Srinivas, MD

Seth A. Strope, MD, MPH

Jonathan Tward, MD, PhD

Przemyslaw Twardowski, MD

Patrick C. Walsh, MD

The NCCN Guidelines are updated at least annually in an evidence-based process integrated with the expert judgment of multidisciplinary panels of expert physicians from NCCN Member Institutions. NCCN depends on the NCCN Guidelines Panel Members to reach decisions objectively, without being influenced or appearing to be influenced by conflicting interests.

All panel member disclosures are available at www.nccn.org.

NCCN 2016 Guideline Prostate Cancer Panel:

James L. Mohler, MD

Andrew J. Armstrong, MD

Robert R. Bahnson, MD

Anthony Victor D'Amico, MD PhD

Brian J. Davis, MD, PhD

James A. Eastham, MD

Charles A. Enke, MD

Thomas A. Farrington,

Celestia S. Higano, MD

Eric Mark Horwitz, MD

Michael Hurwitz, MD, PhD

Christopher J. Kane, MD

Mark H. Kawachi, MD

Michael Kuettel, MD, MBA, PhD

Richard J. Lee, MD, PhD

Joshua J. Meeks, MD, PhD

David F. Penson, MD, MPH

Elizabeth R. Plimack, MD, MS

Julio M. Pow-Sang, MD

David Raben, MD

Sylvia Richey, MD

March Roach, III, MD

Stan Rosenfeld

Edward Schaeffer, MD, PhD

Ted A. Skolarus, MD

Eric J. Small, MD

Guru Sonpavde, MD

Sandy Srinivas, MD

Seth A. Strope, MD, MPH

Johnathon Tward, MD, PhD

All panel member disclosures are available at www.nccn.org.

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guideline:

NCCN Categories of Evidence and Consensus

Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

1c.13 Grade Assigned to the Body of Evidence: NCCN category 1

NCCN 2016 Guideline:

Level of evidence assigned: Category 1

1c.14 Summary of Controversy/Contradictory Evidence: The AUA guideline explains that the use of hormonal therapy may be associated with an increased risk of cardiovascular disease and diabetes and that the use of hormone therapy in men who are at risk for or who are already diagnosed with heart disease and/or diabetes may have their health negatively impacted by the use of hormonal therapy.

Cochrane Review:

The review does not provide a summary of controversy/contradictory evidence. However, the harms studied are referred to in section 1c8.

NCCN 2016 Guideline:

The review does not provide a summary of controversy/contradictory evidence. However, the expected harms are referred to in section 1c8.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Not applicable

Cochrane Review:

Satish K, Shelly M, Harrison C, et al. Neo-adjuvant and adjuvant hormone therapy for localised and locally advanced prostate cancer. Cochrane Database Syst Rev. 2006; (4): CD006019. DOI: 10.1002/14651858.CD006019.pub2. Accessed at: http://www.cochrane.org/CD006019/PROSTATE_neo-adjuvant-and-adjuvant-hormone-therapy-for-localised-and-locally-advanced-prostate-cancer

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

AUA Standard

When counseling patients regarding treatment options, physicians should consider the following:

Based on results of two randomized controlled clinical trials, the use of adjuvant and concurrent hormonal therapy may prolong survival in the patient who has opted for radiotherapy.

High-risk patients who are considering specific treatment options should be informed of findings of recent high-quality clinical trials, including that:

For those considering external beam radiotherapy, use of hormonal therapy combined with conventional radiotherapy may prolong survival

NCCN guideline recommendation

There are several treatment options for patients with high-risk disease. The preferred treatment is 3D-CRT/IMRT with daily IGRT in conjunction with long-term ADT; ADT alone is insufficient. In particular, patients with low volume, high grade tumor warrant aggressive local radiation combined with typically 2-3 years of ADT. [see detailed risk strata below].

<u>Risk Strata: Low, Intermediate, or High –</u>

Low Risk – PSA ≤ 10 mg/dL; AND Gleason score 6 or less; AND clinical stage

T1c or T2a2

Intermediate Risk – PSA > 10 to 20 mg/dL; OR Gleason score 7; OR clinical

stage T2b, and not qualifying for high risk2

High Risk – PSA > 20 mg/dL; OR Gleason score 8 to 10; OR clinically localized

stage T3a1

<u>NCCN 2016 Guidelines:</u>

Men with prostate cancer that is clinical stage T3a, Gleason score 8 to 10, or PSA level greater than 20 ng/mL are categorized by the panel as high risk. Patients with multiple adverse factors may be shifted to the very high-risk category. [See detailed risk strata below]. The preferred treatment is EBRT [external beam radiation therapy] in conjunction with 2 to 3 years of neoadjuvant/concurrent/adjuvant ADT [androgen deprivation therapy] (category 1); ADT alone is insufficient. In particular, patients with low-volume, high-grade tumor warrant aggressive local radiation combined with typically 2 or 3 years of neoadjuvant/concurrent/adjuvant ADT. Fit men in the high-risk group can consider 6 cycles of docetaxel without prednisone after EBRT is completed and while continuing ADT. The combination of EBRT and brachytherapy with our without neoadjuvant/concurrent/adjuvant ADT, is another primary treatment option. However, the optimal duration of ADT in this setting remains unclear. (Category 1) p. 66

Patients at very high risk (locally advanced) are defined by the NCCN Guidelines as men with clinical stages T3b to T4, primary Gleason pattern 5, or more than 4 biopsy cores with Gleason score 8 to 10. The options for this group include: 1) EBRT and long-term ADT (category 1); 2) EBRT plus brachytherapy with or without long-term ADT; 3) EBRT plus ADT and docetaxel; 4) radical prostatectomy plus PLND in selected patients with no fixation to adjunct organs; or 5) ADT for patients not eligible for definitive therapy. (Category 1) p. 66

Risk Strata - Very Low, Low, Intermediate, High, or Very High–

Very Low Risk – PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1c; AND presence of disease in fewer than 3 biopsy cores; AND ≤ 50% prostate cancer involvement in any core; AND PSA density ≤ 0.15 ng/mL/cm3.

Low Risk – PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1 to T2a.

Intermediate Risk – PSA 10 to 20 ng/mL; OR Gleason score 7; OR clinical stage T2b to T2c. Note: patients with multiple adverse factors may be shifted into the high risk category.

High Risk – PSA > 20 ng/mL; OR Gleason score 8 to 10; OR clinically localized stage T3a. Note: Patients with multiple adverse factors may be shifted into the very high risk category.

Very High Risk – Clinical stage T3b to T4; OR primary Gleason pattern 5; OR more than 4 cores with Gleason score 8 to 10.

1c.17 Clinical Practice Guideline Citation: Thompson I, Thrasher JB, Aus G, et al. Guideline for the management of clinically localized prostate cancer: 2007 update. J Urol. 2007;177:2106-2131. Reviewed and validity confirmed 2011.

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011. Available at www.nccn.org

NCCN 2016 Guidelines:

National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 2.2016. Available at www.nccn.org

1c.18 National Guideline Clearinghouse or other URL: http://www.auanet.org/content/clinical-practice-guidelines/clinical-guidelines/main-reports/proscan07/content.pdf and www.nccn.org

Cochrane Review:

Available at http://www.cochrane.org/CD006019/PROSTATE_neo-adjuvant-and-adjuvant-hormone-therapy-for-localised-and-locally-advanced-prostate-cancer

NCCN 2016 Guidelines:

Available at NCCN.org

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: AUA guideline panel members, (specialty): Ian Thompson, M.D., Chair, (Urology) James Brantley Thrasher, M.D., Co-Chair, (Urology) Gunnar Aus, M.D., (Urology) Arthur L. Burnett, M.D., (Sexual Medicine) Edith D. Canby-Hagino, M.D., (Urology) Michael S. Cookson, M.D., (Urology) Anthony V. D'Amico, M.D., Ph.D., (Radiation Oncology) Roger R. Dmochowski, M.D., (Urology) David T. Eton, Ph.D., (Health Services Research) Jeffrey D. Forman, M.D., (Radiation Oncology) S. Larry Goldenberg, O.B.C., M.D., (Urology) Javier Hernandez, M.D., (Urology) Celestia S. Higano, M.D., (Medical Oncology) Stephen R. Kraus, M.D., (Neurourology) Judd W. Moul, M.D., (Urology) Catherine M. Tangen, Dr. P.H., (Biostatistics and Clinical Trials). No disclosures are included in the AUA guideline. NCCN panel member information in section 1c.10.

NCCN 2016 Guideline:

NCCN Prostate Cancer Panel is listed in section 1c.10

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: AUA Guideline Statement Definitions

1. Standard: A guideline statement is a standard if: (1) the health outcomes of the alternative interventions are sufficiently well known to permit meaningful decisions, and (2) there is virtual unanimity about which intervention is preferred.

2. Recommendation: A guideline statement is a recommendation if: (1) the health

outcomes of the alternative interventions are sufficiently well known to permit

meaningful decisions, and (2) an appreciable but not unanimous majority agrees on which intervention is preferred.

3. Option: A guideline statement is an option if: (1) the health outcomes of the

interventions are not sufficiently well known to permit meaningful decisions, or (2) preferences are unknown or equivocal.

NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guidelines:

NCCN Categories of Evidence and Consensus

Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate

1c.23 Grade Assigned to the Recommendation: AUA grade: standard, NCCN grade: category 1

NCCN 2016 Guidelines:

Level of evidence assigned: Category 1

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: Moderate **1c.26** Quality: High**1c.27** Consistency**:** Moderate

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

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

0390_Evidence_MSF5.0_Data-635278494967840120-635932898713011145.doc

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., the benefits or improvements in quality envisioned by use of this measure)*

The use of adjuvant hormonal therapy following external beam radiotherapy is a well-established standard of care for high-risk prostate cancer patients. Multiple large studies have shown that men who receive adjuvant hormonal therapy following external beam radiotherapy can live longer and have a lower risk of recurrence than men who receive radiotherapy alone. In addition, a cost-analysis conducted found that the use of adjuvant hormonal therapy and external beam radiotherapy is cost-effective and adds quality-adjusted life years for patients (1).

Data from several sources indicates that while utilization rates of adjuvant hormonal therapy and external beam radiotherapy have increased, they still remain suboptimal. One study analyzing the CaPSURE database, a provider-based registry, found that the utilization of adjuvant hormonal therapy and external beam radiotherapy for high-risk patients has increased to 80% throughout the past two decades, yet utilization rates have plateaued since 2000 (2). There is rising concern about undertreatment of high-risk prostate cancer patients (3). This suggests greater outreach and education are needed to improve outcomes in care.

Citation:

1. Satish K, Shelly M, Harrison C, et al. Neo-adjuvant and adjuvant hormone therapy for localised and locally advanced prostate cancer. Cochrane Database Syst Rev. 2006; (4): CD006019. DOI: 10.1002/14651858.CD006019.pub2. Accessed at: http://www.cochrane.org/CD006019/PROSTATE_neo-adjuvant-and-adjuvant-hormone-therapy-for-localised-and-locally-advanced-prostate-cancer

2. Cooperberg MR, Janet Cowan, J, Broering JM, et al. High-Risk Prostate Cancer in the United States, 1990-2007. World J Urol. 2008; 26(3): 211–218. doi:10.1007/s00345-008-0250-7.

3. Cooperberg MR, Broering JM, Caroll, PR. Time trends and local variation in primary treatment of localized prostate cancer. J Clin Oncol. 2010;28(7):1117-1123.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI

Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 Registry Performance Rate:

Mean: 93.82%

Minimum: 16.67%

Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:

2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients the last several years are as follows:

2010: 79.60%

2011: 93.50%
2012: 91.10%
2013: 95.40%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 18.70% of eligible professionals reported on Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients for claims and registry. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

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.
An analysis of data from the CaPSURE database, a provider-based registry, conducted by Cooperberg and colleagues found that the utilization of adjuvant hormonal therapy and external beam radiotherapy for high-risk patients has increased to 80% throughout the past two decades, yet utilization rates have plateaued since 2000.

Citation:
1. Cooperberg MR, Janet Cowan, J, Broering JM, et al. High-Risk Prostate Cancer in the United States, 1990-2007. World J Urol. 2008 June ; 26(3): 211–218. doi:10.1007/s00345-008-0250-7.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
While this measure is included in federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
African American/black men have the highest incidence rate of prostate cancer in the United States and are more than twice as likely as white men to die from the disease (1). Between 2008-2012, the average annual prostate cancer incidence rate among African American men was 208.7 cases per 100,000 men, which was 70% higher than the rate in white men. Although prostate cancer incidence and mortality rates have been declining in African American and white men since 1991, the incidence, prevalence, and death rates remain comparably higher among African American men as compared to white men (2).

An analysis of data from the CaPSURE database by Moses and colleagues found significant ethnic and racial differences in the treatment of high-risk prostate cancer. African American men are more likely to receive non-surgical treatment than white men.
White men were 25% less likely than African American men to receive radiation therapy and are 48% less likely to receive hormonal therapy (3).

Citations:
1. National Cancer Institute. Cancer health disparities. http://www.cancer.gov/about-nci/organization/crchd/cancer-health-disparities-fact-sheet. Accessed February 12, 2016.

2. American Cancer Society Cancer Facts and Figures for African Americans 2016-2018. http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-047403.pdf. Accessed February 26, 2016.

3. Moses KA, Paciorek AT, Penson DF, et al. Impact of ethnicity on primary treatment choice and mortality in men with prostate cancer: data from CaPSURE. J Clin Oncol. 2010; 28(6): 1069-1074.

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
220,800 new cases of prostate cancer were diagnosed in 2015 (1). The number of new cases of prostate cancer was 137.9 per 100,000 men per year, based on age-adjusted cases and deaths between 2008-2012. In 2012, there were an estimated 2,795,592 men living with prostate cancer in the United State. Prostate cancer is the third most common type of cancer in the United State and accounts for 13.3% of all new cancer cases. Approximately 14% of men will be diagnosed with prostate cancer at some point in their lifetime, based on 2010-2012 data (2). The annual cost of prostate cancer care in the United States is an estimated $11.85 billion (3).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Siegel RL, Miller KD, Jemal A. Cancer statistics 2015. Ca Cancer J Clin. 2015;65:5-29.

2. SEER Cancer Statistics Factsheets: Prostate Cancer. National Cancer Institute. Bethesda, MD,http://seer.cancer.gov/statfacts/html/prost.html

3. Mariotto AB, Yabroff RK, Shao Y, et al. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst. 2011; 103; 1-12.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not applicable. Not a PRO-PM.

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 subcriteria 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):*
Cancer, Cancer : Prostate

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

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.)*
The measure specifications are included as an attachment with this submission. Additional measure details may be found at http://www.auanet.org/resources/aua-developed-measures.cfm

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:** NQF0390_I9toI10_conversion.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The denominator statement, denominator details, and measure title were revised based on updated risk strata to be consistent with NCCN guidelines.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who were prescribed adjuvant hormonal therapy (GnRH [gonadotropin-releasing hormone] agonist or antagonist)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Once per episode of radiation therapy for all male patients with prostate cancer who receive external beam radiotherapy to the prostate during the 12 month reporting period.

S.6. **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, 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.
Definition:
Prescribed – Includes patients who are currently receiving medication(s) that follow the treatment plan recommended at an encounter during the reporting period, even if the prescription for that medication was ordered prior to the encounter.

For Registry:
To submit the numerator option for patients who were prescribed adjuvant hormonal therapy (GnRH agonist or antagonist), report

the following CPT Category II code:

4164F - Adjuvant (ie, in combination with external beam radiotherapy to the prostate for prostate cancer) hormonal therapy (gonadotropin-releasing hormone [GnRH] agonist or antagonist) prescribed/administered

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All patients, regardless of age, with a diagnosis of prostate cancer at high or very high risk of recurrence receiving external beam radiotherapy to the prostate

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Definitions:
Risk Strata - Very Low, Low, Intermediate, High, or Very High–
Very Low Risk – PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1c; AND presence of disease in fewer than 3 biopsy cores; AND = 50% prostate cancer involvement in any core; AND PSA density = 0.15 ng/mL/cm3.
Low Risk – PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1 to T2a.
Intermediate Risk – PSA 10 to 20 ng/mL; OR Gleason score 7; OR clinical stage T2b to T2c. Note: patients with multiple adverse factors may be shifted into the high risk category.
High Risk – PSA > 20 ng/mL; OR Gleason score 8 to 10; OR clinically localized stage T3a. Note: Patients with multiple adverse factors may be shifted into the very high risk category.
Very High Risk – Clinical stage T3b to T4; OR primary Gleason pattern 5; OR more than 4 cores with Gleason score 8 to 10. (NCCN, 2016)

External beam radiotherapy – external beam radiotherapy refers to 3D conformal radiation therapy (3D-CRT), intensity modulated radiation therapy (IMRT), stereotactic body radiotherapy (SBRT), and proton beam therapy.

Note: Only male patients with prostate cancer with high or very high risk of recurrence will be counted in the performance denominator of this measure.

For Registry:
Any male patient, regardless of age
AND
Diagnosis for prostate cancer (ICD-9-CM): 185
Diagnosis for prostate cancer (ICD-10-CM): C61
AND NOT
Diagnosis for metastatic cancer (ICD-9-CM): 196.0, 196.1, 196.2, 196.3, 196.5, 196.6, 196.8, 196.9, 197.0, 197.1, 197.2, 197.3, 197.4, 197.5, 197.6, 197.7, 197.8, 198.0, 198.1, 198.2, 198.3, 198.4, 198.5, 198.6, 198.7, 198.81, 198.82, 198.89
Diagnosis for metastatic cancer (ICD-10-CM): C77.0, C77.1, C77.2, C77.3, C77.4, C77.5, C77.8, C77.9, C78.00, C78.01, C78.02, C78.1, C78.2, C78.30, C78.39, C78.4, C78.5, C78.6, C78.7, C78.80, C78.89, C79.00, C79.01, C79.02, C79.10, C79.11, C79.19, C79.2, C79.31, C79.32, C79.40, C79.49, C79.51, C79.52, C79.60, C79.61, C79.62, C79.70, C79.71, C79.72, C79.81, C79.82, C79.89, C79.9
AND
Patient encounter during the reporting period (CPT): 77427, 77435
AND
Report the following quality-data code (G-code) to identify the risk of recurrence:
G8465: High or very high risk of recurrence of prostate cancer

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
AUA methodology uses three categories of reasons for which a patient may be excluded from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For this measure, exceptions for not prescribing/administering adjuvant hormonal therapy may include medical reason(s) (eg, salvage therapy) or

patient reason(s). Although this methodology does not require the external reporting of more detailed exception data, the AUA recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The AUA also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement. For example, it is possible for implementers to calculate the percentage of patients that physicians have identified as meeting the criteria for exception. Additional details by data source are as follows:

Documentation of medical reason(s) for not prescribing/administering adjuvant hormonal therapy (eg, salvage therapy)

Documentation of patient reason(s) for not prescribing/administering adjuvant hormonal therapy

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The AUA exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure Adjuvant Hormonal Therapy for High Risk or Very High Risk Prostate Cancer Patients, exceptions may include medical reason(s) (eg, salvage therapy) or patient reason(s) for not prescribing/administering adjuvant hormonal therapy. Although this methodology does not require the external reporting of more detailed exception data, the AUA recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The AUA also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Additional details by data source are as follows:

For Registry:
Documentation of medical reason(s) for not prescribing/administering adjuvant hormonal therapy (eg, salvage therapy)
Append modifier to CPT Category II code: 4164F with 1P

Documentation of patient reason(s) for not prescribing/administering adjuvant hormonal therapy
Append modifier to CPT Category II code: 4164F with 2P

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment or risk stratification

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1) Find the patients who meet the initial patient population (ie, the general group of patients that the performance measure is designed to address).
2) From the patients within the initial patient population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial patient population and denominator are identical.
3) From the patients within the denominator, find the patients who qualify for the Numerator (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator
4) From the patients who did not meet the numerator criteria, determine if the physician has documented that the patient meets any criteria for denominator exception when exceptions have been specified [for this measure: medical reason(s) for not prescribing adjuvant hormonal therapy (eg, salvage therapy) or patient reason(s)]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation.

–Although the exception cases are removed from the denominator population for the performance calculation, the number of patients with valid exceptions should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable. The measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable. The measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Patient eligibility is determined by a set of defined criteria relevant to a particular measure. If data required to determine patient eligibility are missing, those patients/cases would be ineligible for inclusion in the denominator and therefore the patient/case would be deleted.

If data required to determine if a denominator eligible patient qualifies for the numerator (or has a valid exclusion/exception) are missing, this case would represent a quality failure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not applicable. Not a PRO.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Clinician : Team

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinician Office/Clinic, Other
If other: Radiation Oncology Clinic/Department

S.28. **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. The measure is not a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_0390_Adjuvant_Hormonal_Therapy_2016_03_28.docx

NQF #: 0390 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
Five practice sites representing various types, locations and sizes were identified to participate in testing the 3 PCPI/ASTRO/AUA-developed prostate cancer performance measures.
- Site A: hospital, multi-practice sites in urban, rural and suburban settings; 21 physicians; average 9600 oncology/prostate cancer patient visits per month for MD/NP assessment, chemo; submitted PQRS claims for one measure and utilized a full-fledged EHR.
- Site B: physician owned private practice, suburban setting; 4 physicians; average 48 oncology/prostate cancer patients seen per day; submitted PQRS claims for one measure and utilized paper medical records.
- Site C: physician owned private practice, urban setting; 41 physicians; average 2500 oncology/prostate cancer patients seen per month; submitted PQRS claims for two measures and utilized a full-fledged EHR.
- Site D: academic, suburban setting; 9 physicians; average 240 oncology/prostate cancer patients seen per month; submitted PQRS claims for one measure and utilized paper and EHR.
- Site E: academic, urban setting; 14 physicians; average 250 oncology/prostate cancer patients seen per month; collected PQRS data on 3 measures and utilized a full-fledged EHR.

- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**
PCPI Testing Project
Data abstracted from patient records were used to calculate inter-rater reliability for the measure.
91 patient records were reviewed.

Data analysis included:
- Percent agreement; and
- Kappa statistic to adjust for chance agreement.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
PCPI Testing Project
N, % Agreement, Kappa (95% Confidence Interval)
Overall Reliability: 91, 98.9%, 0.972 (0.916-1.000)

Denominator Reliability: 91, 100%, Kappa is noncalculable*
Numerator Reliability: 91, 100%, 0.971 (0.913-1.000)
Exceptions Reliability: 91, 100%, Kappa is noncalculable*

This measure demonstrates almost perfect reliability, as shown in results from the above analysis.

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
The AUA and NCCN guidelines recommend adjuvent hormonal therapy with radiotherapy for high risk prostate cancer patients, for prolonged survival. The measure captures patients receiving external beam radiotherapy in the denominator, and adjuvant hormonal therapy being prescribed in the numerator. Therefore, the evidence directly relates to the specified measure.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.

The panel members are as follows:

Ian Thompson, MD (Co-Chair, urology)
Steven Clauser, PhD (Co-Chair, methodology)
Peter Albertsen, MD (urology)
Colleen Lawton, MD (radiation oncology)
Charles Bennett, MD, PhD, MPP (clinical oncology)
W. Robert Lee, MD, MS, Med (radiation oncology)
Michael Cookson, MD (urology)
Peter A. S. Johnstone, MD, FACR (radiation oncology)
Gregory W. Cotter, MD (radiation oncology)
David F. Penson, MD, MPH (urology)
Theodore L. DeWeese, MD (radiation oncology)
Stephen Permut, MD (family medicine)
Mario Gonzalez, MD (pathology)
Howard Sandler, MD (radiation oncology)
Louis Kavoussi, MD (urology)
Bill Steirman, MA (consumer representative)
Eric A. Klein, MD (urology)
John T. Wei, MD (urology)
Carol Wilhoit, MD (health plan representative)

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
All PCPI performance measures are assessed for content validity by expert Work Group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert Work Group and the measures adjusted as

needed. Other external review groups (i.e. focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity has been quantitatively assessed for this measure. Specifically, the Prostate Cancer Work Group members were asked to empirically assess face validity of the measure. The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.

Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1=Disagree; 3=Neither Disagree nor Agree; 5=Agree

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
The results of the expert panel rating of the validity statement were as follows: N = 14; Mean rating = 4.57.

Percentage in the top two categories (4 and 5): 92.86%

Frequency Distribution of Ratings
1 – 0
2 – 0
3 – 1
4 – 4
5 – 9

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
- 91 patient records were reviewed for this measure.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Exceptions were analyzed for frequency and variability across providers.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
PCPI Testing Project
N, % Agreement, Kappa (95% Confidence Interval)
Exceptions Reliability: 91, 100%, Kappa is noncalculable*

This measure demonstrates perfect reliability, as shown in results from the above analysis.

The exception rate for this measure is 3.3%

*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
This measure is not risk adjusted.

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
This measure is not risk adjusted.

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
Not applicable.

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:** As a process measure, no risk adjustment is necessary.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
- 91 patient records were reviewed for this measure.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

CMS Physician Quality Reporting Initiative:
Clinical Condition and Measure: #104
2,736 patients were reported on for the 2008 program, the most recent year for which data are available.

In 2009 the following was reported for this measure:
Eligible Professionals: 4,114
Professionals Reporting >=1 Valid QDC: 431
% Professionals Reporting >=1 Valid QDC: 10.48%
# Professionals Satisfactorily Reporting: 101
% Professionals Satisfactorily Reporting: 23.43%

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
PCPI Testing Project
Data analysis performed on the measure included:
Average measure performance rate overall and by site, performance rate range by site and overall standard deviation for the measure.

CMS Physician Quality Reporting Initiative:
The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PCPI Testing Project
Measure rate without exceptions: N= 91 Mean = 75.8% Standard Deviation= 0.4305
The performance rate by site is as follows, where n is the number of performance events by site:

A 0.6670 n=21
B 0.8890 n=9
D 0.8000 n=30
E 0.7420 n=31

The performance rate range is .2220. Although this study captured performance on 91 events, the data were not captured at the physician level, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a meaningful calculation of variation in performance across organizations due to the small sample size of sites (n=4) in this study.

CMS Physician Quality Reporting Initiative
This measure was used in the 2008-2011 CMS Physician Quality Reporting Initiative Claims and Registry options and group reporting option available in 2011.

There is a gap in care as shown by this 2008 data, the only year for which distribution by quartile/decile is available.

83.41% of patients reported on did not meet the measure.

10th percentile: 0.00%
25th percentile: 0.00%
50th percentile: 7.69%
75th percentile: 22.22%
90th percentile: 50.00%

The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 22.22, and indicates that 50% of physicians have performance on this measure ranging from 0.00% and 22.22%. A quarter of reporting physicians have performance on this measure which is greater than 22.22%, while a quarter have performance on this measure at 0.00%.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
PCPI Testing Project
- 39 Medicare patient records of the 91 patient records were reviewed.
- The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010.
- Chart abstraction was performed between 8/8/2011 and 11/3/2011.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**
PCPI Testing Project

Parallel forms reliability testing was performed. PQRS claims were reviewed and compared to a manual review of claims information.

Data analysis included:
- Percent agreement

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**
PCPI Testing Project
N, % Agreement
39, 100%

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)***:** We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one's ancestry) and language need (a rating of spoken English language proficiency of less than very well and one's preferred language for health-related encounters)."(2)

References:
(1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.

(2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? (*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

New reliability testing, empirical validity testing of the measure score, face validity.

Measure Number (*if previously endorsed*)**:** 390

Measure Title: Adjuvant Hormonal Therapy for High Risk Prostate Cancer Patients

Date of Submission: 3/11/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.
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.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 data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

1.3. What are the dates of the data used in testing? The data are for the time period January 2014 through December 2014 and cover the entire United States.

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> 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: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> 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 total number of physicians reporting on this measure, via the registry option, in 2014, is 86. Of those, 20 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 634 quality events. For this measure, 23.3 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 21.5 for the remaining 430 events. The range of quality reporting events for 20 physicians included is from 65 to 10. The average number of quality reporting events for the remaining 76.7 percent of physicians that aren't included is 1.1.

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

There were 430 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure and remained after exceptions were removed.

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.

The same data sample was used for reliability testing and exceptions analysis.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient-level socio-demographic (SDS) variables were not captured as part of the testing for this measure.

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 of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

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*)

This measure has 0.73 reliability when evaluated at the minimum level of quality reporting events and 0.85 reliability at the average number of quality events.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is high.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

---

To satisfy NQF's ICD-10 Conversion Requirements, we are providing the information below:

- NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM
  Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.
- NQF ICD-10-CM Requirement 2: Coding Table
  See attachment in S.2b
- NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes
  The PCPI's ICD-10 conversion approach was used to identify ICD-10 codes for this measure. The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

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

The expert panel included 21 members. Panel members were comprised of experts from the AUA Quality Improvement and Patient Safety Committee. The list of expert panel members is as follows:

Christopher Tessier, MD (Chair)
Timothy Averch, MD (Vice Chair)
Daniel Barocas, MD
Kristin Chrouser, MD, MPH
Machele Donat, MD
John L. Gore, MD
Fernando Kim, MD
Danil V. Makarov, MD
Jodi Maranchie, MD
Matthew Nielsen, MD
Caleb Nelson, MD
Elliot Paul, MD
John Stoffel, MD
J. Quentin Clemens, MD, MSCI
Roger Dmochowski, MD
Deborah Lightner, MD
Christopher Saigal, MD
J.Stuart Wolf, Jr., MD

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

The results of the expert panel rating of the validity statement were as follows: N = 15; Mean rating = 4.5 and 100% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)
2 – 0 responses
3 – 0 responses (Neither Agree nor Disagree)
4 – 7 responses
5 – 8 responses (Strongly Agree)

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

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

Exceptions include:

- Documentation of medical reason(s) for not prescribing adjuvant hormonal therapy (eg, salvage therapy)
- Documentation of patient reason(s) for not prescribing adjuvant hormonal therapy

Exceptions were analyzed for frequency across providers.

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

Exceptions Analysis:

Amongst the 20 physicians with the minimum (10) number of quality reporting events, there were a total of 204 exceptions reported. The average number of exceptions per physician in this sample is 10.2. The overall exception rate is 32.2%.

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

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons.

Without these being removed, the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives.

AUA recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. AUA also advocates for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Not applicable

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

Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

Not applicable

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

Not applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

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

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

Measures of central tendency, variability, and dispersion were calculated.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
Based on the sample of 20 included physicians, the mean performance rate is 0.93 the median performance rate is 1.00 and the mode is 1.00. The standard deviation is 0.15. The range of the performance rate is 0.57, with a minimum rate of 0.43 and a maximum rate of 1.00. The interquartile range is 0.06 (.94-1.00).

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

The range of performance from .43 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

This test was not performed for this measure.

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

This test was not performed for this measure.

2b6.3. 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)

This test was not performed for this measure.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Data are not available to complete this testing.

2b7.2. 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*)

Data are not available to complete this testing.

2b7.3. 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*)

Data are not available to complete this testing.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

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

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

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.**

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.

The AUAER and PCPI jointly hold all rights on the Measure, including copyright, in perpetuity, in all forms and media throughout the world, for the full term of copyright, including renewals. Notwithstanding the foregoing, the Measure will be available to the public free of charge for use in non-commercial endeavors without seeking any permissions (notices on the measures provide this permission, e.g., use by health care providers in connection with their practices is not a commercial use).

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting PQRS http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/pqrs/index.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) AQUA Registry http://www.auanet.org/resources/aqua.cfm Michigan Urological Surgery Improvement Collaboration (MUSIC) http://musicurology.com/

4a.1. For each CURRENT use, checked above, provide:

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

1. Physician Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)

Purpose: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013. Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented a phased approach to public reporting performance information on the Physician Compare Web site. CMS announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in late 2017.

2. AQUA Registry – Sponsored by the American Urological Association

Purpose: As part of its ongoing commitment to improving the quality of care for patients with urologic disease, the American Urological Association launched the AUA Quality (AQUA) Registry with a pilot project in June 2014. The AQUA Registry is a national urologic disease registry designed to measure and report healthcare quality and patient outcomes. Through the aggregation and organization of both clinical- and patient-reported data on diagnostic and therapeutic interventions, outcomes and resource utilization, the Registry will provide the urologic community with a definitive resource for informing and advancing urology within the United States. The AQUA Registry currently focuses on prostate cancer, but it will gradually expand to include other urological conditions when it becomes a Qualified Clinical Data Registry (QCDR). It is anticipated that the registry will achieve QCDR status in early 2016. The registry is currently in the early stages of data aggregation and has plans for continued expansion in 2016 and

beyond. Beginning in late 2016, CMS will require QCDRs to publicly report on PQRS and non-PQRS measure data at the individual EP level following the first year of reporting by the QCDR.

3. Michigan Urological Surgery Improvement Collaboration (MUSIC)-- Sponsored by Blue Cross and Blue Shield of Michigan as part of the BCBSM Value Partnerships program

Purpose: The overall aims of the collaborative include, among others, evaluating and improving patterns of care in the radiographic staging of men with newly diagnosed prostate cancer, reducing biopsy-related complications and assessing repeat biopsy patterns, improving patient outcomes after radical prostatectomy, enhancing patient-centered decision making among men considering local therapy for early-stage prostate cancer, and understanding and reducing variation in the use of androgen deprivation therapy. Participating practices number 45 and submit data to a clinical registry maintained by the MUSIC Coordinating Center and tri-annual consortium-wide meetings are held each year to discuss data, review risk-adjusted measures of processes of care and patient outcomes, and identify strategies and best practices for quality improvement.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI
Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 Registry Performance Rate:
Mean: 93.82%
Minimum: 16.67%
Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients the last several years are as follows:
2010: 79.60%
2011: 93.50%
2012: 91.10%
2013: 95.40%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 18.70% of eligible professionals reported on Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients for claims and registry. As a result, performance rates may not

be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

4b.2. **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.**
While the AUA creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1. Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences at this time, but we take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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)**

0220 : Adjuvant hormonal therapy
0389 : Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients
1853 : Radical Prostatectomy Pathology Reporting

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

5a. Harmonization

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 completely harmonized?

No

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

For measure 0220, Adjuvant Hormonal Therapy, the related measure focuses on adjuvant hormonal therapy for breast cancer patients, which is not consistent with the target population addressed in measure 0390. While this is the same action, it is a different drug and target population addressed in each measure. The related measure 0389, Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients addresses the use of bone scan in low-risk prostate cancer patients which is a different quality action from measure 0390. The two measures do not share similar target populations and address different aspects of prostate cancer care. The related measure 1853, Radical Prostatectomy Pathology Reporting, addresses the percentage of radical prostatectomy pathology reports that include the pT category, the pN category, the Gleason score and a statement about margin status, which is a different action than measure 0390. The two measures do not share similar target populations and address different aspects of prostate cancer care.

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):** American Urological Association
Co.2 **Point of Contact:** Suzanne, Pope, spope@auanet.org, 410-689-4026-
Co.3 **Measure Developer if different from Measure Steward:** American Urological Association
Co.4 **Point of Contact:** Suzanne, Pope, spope@auanet.org, 410-464-4904-

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.
Timothy D. Averch, MD, FACS (urology)
John L. Gore, MD, MS, FACS (urology)
Christopher D. Tessier, MD (urology)
Ian Thompson, MD (Co-Chair, urology)
Steven Clauser, PhD (Co-Chair, methodology)
Peter Albertsen, MD (urology)
Colleen Lawton, MD (radiation oncology)
Charles Bennett, MD, PhD, MPP (clinical oncology)
W. Robert Lee, MD, MS, Med (radiation oncology)
Michael Cookson, MD (urology)
Peter A. S. Johnstone, MD, FACR (radiation oncology)
Gregory W. Cotter, MD (radiation oncology)
David F. Penson, MD, MPH (urology)
Theodore L. DeWeese, MD (radiation oncology)
Stephen Permut, MD (family medicine)
Mario Gonzalez, MD (pathology)
Howard Sandler, MD (radiation oncology)
Louis Kavoussi, MD (urology)
Bill Steirman, MA (consumer representative)
Eric A. Klein, MD (urology)
John T. Wei, MD (urology)
Carol Wilhoit, MD (health plan representative)

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 09, 2015
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?** 09, 2017

Ad.6 **Copyright statement:** The AUAER and PCPI shall jointly hold all rights on the Measure, including copyright, in perpetuity, in all forms and media throughout the world, for the full term of copyright, including renewals. Notwithstanding the foregoing, the Measure will be available to the public free of charge for use in non-commercial endeavors without seeking any permissions (notices on the measures provide this permission, e.g., use by health care providers in connection with their practices is not a commercial use).
Ad.7 **Disclaimers:** Please see the copyright statement above in AD.6 for disclaimer information.

Ad.8 **Additional Information/Comments:**

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 #: 0459
Measure Title: Risk-Adjusted Length of Stay >14 Days after Elective Lobectomy for Lung Cancer
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: Percentage of patients aged 18 years and order undergoing elective lobectomy for lung cancer who had a prolonged length of stay >14 days
Developer Rationale: It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Prolonged length of stay after pulmonary lobectomy is both a surrogate marker of morbidity, but also, importantly, a direct marker of increased resource utilization. Knowing their rate of risk-adjusted prolonged length of stay gives lower performing thoracic programs the opportunity to design quality improvement initiatives. These should lead to better patient outcomes and decreased resource utilization.

Numerator Statement: Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer who had a prolonged length of stay >14 days
Denominator Statement: Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer
Denominator Exclusions: None

Measure Type: Outcome
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Jul 31, 2008

Maintenance of Endorsement -- Preliminary Analysis

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 outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer stated that prolonged length of stay after pulmonary lobectomy is both a surrogate marker of morbidity, but also, importantly, a direct marker of increased resource utilization. Lower performing thoracic programs have the opportunity to design quality improvement initiatives when they know their rate of risk

adjusted prolonged length of stay – this should lead to better patient outcomes and decreased resource utilization.

- The developer noted several factors that may impact lobectomy outcomes such as age, gender, payer, comorbidities, surgical approach, good patient selection, and implementation of a multimodal enhanced recovery pathway.

Guidance from the Evidence Algorithm: Health outcome measure→The relationship between the outcome and at least one process is identified and supported by the stated rationale→Pass

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

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 the following performance data from the STS General Thoracic Surgery Database (GTSD) for patients that underwent elective lobectomy for lung cancer between July 1, 2012 and June 30, 2015.

	Total	Midwest	Northeast	South	West
# of participants	244	49	73	82	40
# of patients	23174	5020	7756	7402	2996

Participant-specific PLOS risk adjusted rates (RAR) (%)

Mean	4.3	4.2	4.4	4.4	4.1
SD	1.2	1.1	1.1	1.4	1.0
IQR	1.2	1.4	1.2	1.2	1.1
Minimum	1.9	2.1	2.3	2.2	1.9
10% percentile	3.0	2.7	3.2	3.1	2.9
20% percentile	3.4	3.3	3.6	3.4	3.4
30% percentile	3.7	3.5	3.8	3.6	3.6
40% percentile	3.9	3.9	4.0	3.9	3.9
Median	4.2	4.1	4.2	4.2	4.0
60% percentile	4.4	4.4	4.5	4.3	4.1
70% percentile	4.6	4.6	4.7	4.5	4.5
80% percentile	4.9	5.0	5.1	4.9	4.7
90% percentile	5.7	5.6	5.8	6.6	5.4
Maximum	10.4	7.4	7.8	10.4	6.8

- The developer stated that prolonged length of stay (PLOS) occurred in **4.1%** (948/23,174) of eligible patients reviewed from July 1, 2012 through June 30, 2015.
- For endorsement maintenance, NQF asks for performance scores on the measure as specified, *current and over time*. The developer did not provide performance data on the measure from 2008 (when first endorsed) through 2012.

Disparities:

- The developer did not provide data on disparities from the measure as specified – this is encouraged for endorsement maintenance.

Questions for the Committee:

- *Without performance data prior to 2012, is it possible to determine if there is a gap in care that warrants a*

national performance measure?

- *Are you aware of any disparities data that exists 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 to Support Measure Focus

Comments:

Among the options for a provider wishing to potentially improve one's measured performance is selection of patients and surgical approach. Recovery pathways may reduce duration of stay and/or impact resource use.

1b. Performance Gap

Comments:

The 2012-2015 analysis seems to indicate that the range of PLOS risk adjusted rates is greater than originally reported by Wright et al. in 2008, with the lower end showing increased range.
Is 14 days still an appropriate threshold for defining PLOS?
I'm unaware of documented disparities among population subgroups, but the risk adjustment and an analysis provided by the developer suggest that age and gender are relevant.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Electronic Clinical Data : Registry

Specifications:

- The measure title was changed from Risk-Adjusted Morbidity: Length of Stay >14 Days after Elective Lobectomy for Lung Cancer to Risk-Adjusted Length of Stay >14 Days after Elective Lobectomy for Lung Cancer.
- This is a clinician-level measure.
- The measure is risk adjusted.
- The numerator of this measure is: Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer who had a prolonged length of stay >14 days.
- The denominator time window was changed from 12 months to 36 months. The 36-month time window is necessary to obtain appropriate sample sizes for this measure.
- The denominator is: Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer
- There are no exclusions.
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The developer refers to numerator and denominator sections for detailed information about the calculation algorithm.
- The developer addresses how missing data are handled.
- STS General Thoracic Surgery Database (GTSD) is the registry identified as the specific data source for this measure.
- Collection instrument available at measure specific web page.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure is consistently implemented?*

2a2. Reliability Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- In the prior submission, the developer assessed test-retest reliability by comparing the results of estimated hospital rates of prolonged stay between two consecutive 6-month time intervals during 2009. The Pearson correlation between hospital-specific rates of prolonged stay in the 1st vs. 2nd half of 2009 was 0.31.

Describe any updates to testing

- The developer provided updated reliability testing of the measure score – see below

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 23,174 operations from 244 STS General Thoracic Surgery Database participants from July 1, 2012 to June 30, 2015 (36 months).
- The developers conducted a signal-to-noise analysis using the Pearson correlation coefficient to test the measure score reliability; this is an appropriate method. The signal 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.
    - Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Guide from Evans (1996) suggests for the absolute value of r: 0.00-0.19 as "very weak", 0.20-.039 "weak",0.40-0.59 "moderate", 0.60-0.79 "strong", and 0.80-1.0 "very strong". [Evans, J.D. (1996) Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove.]

Results of reliability testing

- The developers provided the reliability results below and noted that the reliability of the measure score increased as the volume of minimum procedures per year for participants increased .

	All participants	≥10 procedures per year	≥20 procedures per year	≥30 procedures per year	≥40 procedures per year
Number of participants	244	184	132	95	72
Reliability	37.6%	44.5%	49.5%	56.1%	63.8%
95% CrI for reliability	(26.1%, 48.9%)	(32.3%, 55.5%)	(36.2%, 61.5%)	(41.3%, 68.6%)	(48.6%, 76.4%)

Guidance from the Reliability Algorithm: Precise specifications (Box 1)→empiric reliability testing (Box 2) → performance measure score (Box 4) → signal-to-noise analysis used to calculate reliability rates (Box 5) → moderate certainty/confidence that performance measure scores are reliable (Box6b) →Moderate

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is it likely this measure is consistently implemented?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- The developers previously stated that beginning in the fall of 2010 they would conduct patient-level data element validity testing . Twenty randomly selected lobectomy cases previously submitted to the STS data warehouse would be chosen for review of 30 individual data elements. The developer also stated that validity was confirmed by an expert panel of thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Task Force on Quality Initiatives and the STS Workforce on National Databases. No testing results were provided.

Describe any updates to validity testing: see empirical validity testing below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**
☐ **Empirical validity testing of the measure score**

Validity testing method:

- The dataset used for data element validity testing included 10% of randomly selected STS GTSD participants from 2013 to 2015: N = 25 (2015); N = 24 (2014); N = 18 (2013). Twenty cases (at least 15 lobectomy and up to 5 esophagectomy) that were previously submitted to the STS data warehouse were re-abstracted and compared to the 'gold standard'; this is an appropriate method. Agreement rates were calculated for 40 STS GTSD V2.2 individual data elements.
- The developer provided data on the relationship between process variables and agreement rate, though, this does not meet NQF validity testing requirements.
- The developer also stated that the measure is regarded as useful and valid by its intended users and differences in the measure across participants are clinically meaningful; however, the information provided is not sufficient to meet NQF's requirement for face validity.

Validity testing results:

- The developer stated that in 2015, there were 14,854 total variables abstracted and of those 14,412 variables matched. Individual data elements were included in the following categories: pre-operative evaluation, diagnosis and procedures, post-operative events, and discharge.
- Agreement rates for the individual data elements ranged from 84.15% (diabetes control) to 100.0% (esophageal

cancer, date of surgery, gastric outlet, and discharge date).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer did not identify any exclusions

Questions for the Committee:

- *Should exclusions be identified for this measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary:

Description of the model

- The measure is risk adjusted using a logistic regression model. The developers stated that covariates were selected based on clinical relevance, literature review, and empirical analysis conducted by a panel of physicians and statisticians.
- The developers did not provide the details of the statistical methods and criteria used to select patient factors in the risk model, though they state that the details of the model development were published in 2008 (Wright et al). The final model includes 10 variables.
- The developer used the C-index and Hosmer-Lemeshow Goodness-of-Fit statistical methods to assess model discrimination calibration but did not provide a detailed description of the analyses.
  - The C-index or c-statistic, reflects how accurately a statistical model is able to distinguish between a patient with an outcome and a patient without an outcome. C-statistic values can range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of patients with and without the outcome of interest and a value of 1.0 indicates that the model perfectly identifies those with and without the outcome of interest. Generally, a c-statistic of at least 0.70 is considered acceptable.
  - The Hosmer-Lemeshow test is used to determine the goodness of fit of the logistic regression model.

Performance of the model

- The developer reported the following statistical results:
  - C-statistic: **0.672**
  - Hosmer and Lemeshow Goodness-of-Fit Test: p-value=**0.94** (Chi-Square=**2.89**, df=**8**)
  - The developer did not provide risk decile plots.

SDS Conceptual Description

- The developer performed a literature search to help inform their conceptualization of the pathways by which SDS factors affect prolonged length of stay after lobectomy in the acute care setting. The developers stated that there is very little written in the literature about sociodemographic (SDS) factors related to length of stay after lobectomy but noted a couple of studies that considered the following SDS factors:
  - Insurance status (Medicaid vs. private insurance)
  - Rural, low-volume hospitals, Medicaid or lower median income patients underwent VATS lobectomy less often; VATS lobectomy was associated with shorter hospital stays.
- The developer stated that given the lack of consistent, compelling evidence regarding SDS factors and length of stay, they are **not** included in the current measure being submitted.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*

○ *Are all of the risk adjustment variables present at the start of care?*

○ *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Participant-specific risk adjusted rates (RAR) for the prolonged length of stay after lobectomy were estimated within a Bayesian hierarchical logistic regression model.
- Participant-specific RARs were plotted along with corresponding 95% credible intervals to illustrate the between-participant variation.
- The identified differences in performance are clinically meaningful and allow differentiation between participants based on estimates and 95% credible intervals.

Question for the Committee:

○ *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- N/A

2b7. Missing Data

- The developer managed the missing data with imputation. Missing %FEV1 values were imputed utilizing median of the observed %FEV1 values. For binary risk factors, missing values were considered as indicating absence of the risk factor.
- Variables with missing data were: renal dysfunction (4.49%), induction therapy (2.41%), %FEV1 (4.28%), smoking status (0.02%). Other covariates had no missing data.
- The missing data approach was compared, as a sensitivity analysis, to multiple imputations and the results were insensitive to the approach.

Guidance from Validity Algorithm: Precise specifications (Box 1)→ potential threats to validity assessed (Box 2)→ empirical validity testing (Box 3)→validity testing conducted with patient-level data elements (Box 10) →Data element validity compared to gold standard for 40 individual data elements (Box 11)→High/Moderate certainty or confidence that data used in the measure are valid (Box 12a) →Moderate (Moderate is highest eligible rating)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Will using a 3 year time window for measurement if the time-period for comparison is shorter complicate use of the measure (i.e. if one is below average one year but then dramatically improves might one continue to be regarded as below average for a couple years until the "bad" year moves out of the 3 year window)?

The data elements are well defined except for age, which is lumped in 10 year increments in a manner that isn't obvious. Appropriate codes are included. The calculation is clearly described, although it isn't clear why year of surgery is included as a predictor.

It's likely to be implemented uniformly.

**Given the 2010 article by Paul et al., it appears that surgical approach may help predict of PLOS. This isn't included in the calculation. On one hand, if it is a predictor, not having it in the model might advantage thoracoscopy and maybe that's good if that is an intent; however, it may make it hard to distinguish someone who does excellent thoracotomies from someone with terrible thoracoscopy skills. Different surgical approaches almost seems like defining two different groups of patients. **

**Is 14 days still an appropriate threshold for defining PLOS? **

**Is PLOS the same in the context VATS and open lobectomies? **

2a2. Reliability Testing

Comments:

**If the PLOS as defined is an appropriate measure, then it appears able to fairly reliably and consistently identify differences among surgeons. **

2b2. Validity Testing

Comments:

**Gathering the data and scoring seem achievable and consistent such that you could feel confident drawing some conclusions about quality when scores are substantially different between two providers. **

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

**I'd like to understand how/why year of surgery factors in. **

Risk adjustment seems reasonable; there is not compelling evidence that additional SDS factors would improve the validity.

I wouldn't expect any substantial problem from occasional missing data.

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.
- All data elements from participating institutions are submitted to the STS GTSD in an electronic format following a standard set of data specifications.
- STS GTSD participating institutions utilize data entry software products that are approved for the purposes of collecting and submitting STS GTSD data elements.
- There are no direct costs to collect data for this measure. STS General Thoracic Surgery Database participants (single surgeon or a group of surgeons) pay annual fees of $550 per surgeon for STS members and $700 per surgeon for non-STS members. In addition, there is a cost associated with purchasing data collection software which varies across vendors. STS GTSD participants have a separate agreement with their vendor, a process in which STS is not involved.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

**The data elements are likely to be routinely accessible, sometimes electronically as discrete data elements, but probably they'll often need to be abstracted for submission. **

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details :

- In 2017, STS is planning to launch the general thoracic surgery component of STS Public Reporting Online. This is currently in the planning stage. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online
- The STS National Database is a Qualified Clinical Data Registry (QCDR). Prolonged length of stay is one of many measures that STS reports to CMS on behalf of consenting STS Adult Cardiac Surgery Database surgeons. For the STS General Thoracic Surgery Database (GTSD), physician quality reporting is in the planning stage.
- STS GTSD leaders and STS staff are reviewing general thoracic measures for inclusion in physician quality reporting.
- STS intends to include general thoracic measures in its 2017 QCDR self-nomination.

Improvement results:

- The developer provided the information below. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

	Total	Midwest	Northeast	South	West
# of participants	244 (100%)	49 (20.1%)	73 (29.9%)	82 (33.6%)	40 (16.4%)
# of patients	23174 (100%)	5020 (21.7%)	7756 (33.5%)	7402 (31.9%)	2996 (12.9%)

Unexpected findings (positive or negative) during implementation :

- The developer reports no additional difficulties or unexpected findings or benefits, apart from those included throughout the submission form.

Potential harms:

- The developer reports no unintended consequence were noted.

Feedback :

- Measure has not been reviewed by MAP.

***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

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
**Looks like there is a plan for public reporting. I'm not sure how prominent the proposed reporting program is or if it would guide many patients, but inclusion in an accountability program would be beneficial. Seems likely that low performance would cause a provider to first reassess selection of patients and surgical procedure. **

Criterion 5: Related and Competing Measures
Related or competing measures N/A **Harmonization** N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0459

Measure Title: Risk-Adjusted Length of Stay >14 Days after Elective Lobectomy for Lung Cancer

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

Date of Submission: 3/14/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering 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:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

☒ Health outcome: prolonged length of stay

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Prolonged length of stay after pulmonary lobectomy is both a surrogate marker of morbidity, but also, importantly, a direct marker of increased resource utilization. Knowing their rate of risk adjusted prolonged length of stay gives lower performing thoracic programs the opportunity to design quality improvement initiatives. These should lead to better patient outcomes and decreased resource utilization.

Lobectomy is considered the gold standard for patients with stage I lung cancer and is the most common operation performed for early stage lung cancer within the STS database (1,2). Approximately 50,000 lobectomies are performed in the United States each year. Prolonged length of stay after lobectomy has been demonstrated to be a surrogate marker for complications after lobectomy (3). Patients within the STS GTSD with a prolonged length of stay >14 days have a markedly increased mortality rate, 10.8% vs. 0.7% ($p<0.0001$). Patients with lengths of stay longer than 16 days have increased 30-day readmission rates as well (4). Both of these intuitively demonstrate the increase in resource utilization from prolonged length of stay. Variability exists in length of stay following even uncomplicated lobectomy and can be driven by factors such as age, gender, payer, comorbidities, and surgical approach (5). Reducing postoperative complications and length of stay requires good patient selection and may be enhanced by cardiopulmonary exercise testing in selected patients with marginal pulmonary function (6). Surgeon factors, such as surgical approach chosen for lobectomy, are a strong predictor of postoperative complications and length of stay. A propensity matched analysis within the STS GTSD demonstrated that thoracoscopic lobectomy reduced both complications as well as length of stay when compared with open lobectomy (7). Similarly, an analysis of the National Inpatient Sample demonstrated fewer in-hospital complications and shorter length of stay with a thoracoscopic

approach and also demonstrated that only 15% of lobectomies were being done thoracoscopically (8). In-hospital, post-surgical pathways can also alter prolonged length of stay. Implementation of a multimodal enhanced recovery pathway has been associated with a decrease in length of stay and complications with no increase in readmissions (9).

1. Ginsburg RJ, Rubinstein LV. Randomized trial of lobectomy versus limited resection for T1N0 non-small cell lung cancer. Lung Cancer Study Group. Ann Thorac Surg 1995;60:615-22.
2. Boffa DJ, Allen MS, Grab JD, Gaissert HA, Harpole DH, Wright CD. Data from The Society of Thoracic Surgeons General Thoracic Surgery database: the surgical management of primary lung tumors. J Thorac Cardiovasc Surg. 2008 Feb;135(2):247-54.
3. Wright CD, Gaissert HA, Grab JD, O'Brien SM , Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.
4. Freeman RK, Dilts JR Ascioti AJ et al. A comparison of length of stay, readmission rate, and facility reimbursement after lobectomy of the lung. Ann Thorac Surg 2013;96(5):1740-5
5. Giambrone GP, Smith MC, Wu X, et al. Variability in length of stay after uncomplicated pulmonary lobectomy: Is length of stay a quality metric or a patient metric. Eur J Cardiothorac Surg 2016;Jan 27: Epub ahead of print.
6. Brunelli A, Belardinelli R, Pompii C, et al. Minute ventilation to carbon dioxide output (VE/VCO2) slope is the strongest predictor of respiratory complications and death after pulmonary resection. Ann Thorac Surg 2012;93(6):1802-6.
7. Paul S, Altorki NK, Sheng S, et al. Thoracoscopic lobectomy is associated with lower morbidity than open lobectomy: a propensity-matched analysis from the STS database. J Thoracic Cardiovasc Surg 2010;139(2):366-78.
8. Paul S, Sedrakyan A, Chiu YL, et al. Outcomes after lobectomy using thoracoscopy vs thoracotomy: a comparative effectiveness analysis utilizing the National Inpatient Sample. Eur J Cardiothorac Surg 2013;43(4):813-7.
9. Madani A, Fiore JF Jr, Wang Y, et al. An enhanced recovery pathway reduces duration of stay and complications after open pulmonary lobectomy. Surgery 2015;158(4):899-908.

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

See response in 1a.2.

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

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

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

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

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

☐ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

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

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

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

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

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

☐ Yes → *complete section 1a.7*

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

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

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

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

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

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

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

Complete section 1a.7

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

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

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

Complete section 1a.7

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

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

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

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

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

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

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

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

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

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

1a.8 OTHER SOURCE OF EVIDENCE

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

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

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

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1a._Evidence_-_0459_Lobectomy_LOS.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Prolonged length of stay after pulmonary lobectomy is both a surrogate marker of morbidity, but also, importantly, a direct marker of increased resource utilization. Knowing their rate of risk-adjusted prolonged length of stay gives lower performing thoracic programs the opportunity to design quality improvement initiatives. These should lead to better patient outcomes and decreased resource utilization.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see the Appendix

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Please see the Appendix

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Lobectomy is considered the gold standard for patients with stage I lung cancer and is the most common operation performed for early stage lung cancer within the STS database (1,2). Approximately 50,000 lobectomies are performed in the United States each year. Prolonged length of stay after lobectomy has been demonstrated to be a surrogate marker for complications after lobectomy (3). Patients within the STS GTSD with a prolonged length of stay >14 days have a markedly increased mortality rate, 10.8% vs. 0.7% ($p<0.0001$). Patients with lengths of stay longer than 16 days have increased 30-day readmission rates as well (4). Both of these intuitively demonstrate the increase in resource utilization from prolonged length of stay. Variability exists in length of stay following even uncomplicated lobectomy and can be driven by factors such as age, gender, payer, comorbidities, and surgical approach (5). Reducing postoperative complications and length of stay requires good patient selection and may be enhanced by cardiopulmonary exercise testing in selected patients with marginal pulmonary function (6). Surgeon factors, such as surgical approach chosen for lobectomy, are a strong predictor of postoperative complications and length of stay. A propensity matched analysis within the STS GTSD demonstrated that thoracoscopic lobectomy reduced both complications as well as length of stay when compared with open lobectomy (7). Similarly, an analysis of the National Inpatient Sample demonstrated fewer in-hospital complications and shorter length of stay with a thoracoscopic approach and also demonstrated that only 15% of lobectomies were being done thoracoscopically (8). In-hospital, post-surgical pathways can also alter prolonged length of stay. Implementation of a multimodal enhanced recovery pathway has been associated with a decrease in length of stay and complications with no increase in readmissions (9).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Ginsburg RJ, Rubinstein LV. Randomized trial of lobectomy versus limited resection for T1N0 non-small cell lung cancer. Lung Cancer Study Group. Ann Thorac Surg 1995;60:615-22.
2. Boffa DJ, Allen MS, Grab JD, Gaissert HA, Harpole DH, Wright CD. Data from The Society of Thoracic Surgeons General Thoracic Surgery database: the surgical management of primary lung tumors. J Thorac Cardiovasc Surg. 2008 Feb;135(2):247-54.
3. Wright CD, Gaissert HA, Grab JD, O´Brien SM , Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.
4. Freeman RK, Dilts JR Ascioti AJ et al. A comparison of length of stay, readmission rate, and facility reimbursement after lobectomy of the lung. Ann Thorac Surg 2013;96(5):1740-5
5. Giambrone GP, Smith MC, Wu X, et al. Variability in length of stay after uncomplicated pulmonary lobectomy: Is length of stay a quality metric or a patient metric. Eur J Cardiothorac Surg 2016;Jan 27: Epub ahead of print.
6. Brunelli A, Belardinelli R, Pompii C, et al. Minute ventilation to carbon dioxide output (VE/VCO2) slope is the strongest predictor of respiratory complications and death after pulmonary resection. Ann Thorac Surg 2012;93(6):1802-6.
7. Paul S, Altorki NK, Sheng S, et al. Thoracoscopic lobectomy is associated with lower morbidity than open lobectomy: a propensity-matched analysis from the STS database. J Thoracic Cardiovasc Surg 2010;139(2):366-78.
8. Paul S, Sedrakyan A, Chiu YL, et al. Outcomes after lobectomy using thoracoscopy vs thoracotomy: a comparative effectiveness analysis utilizing the National Inpatient Sample. Eur J Cardiothorac Surg 2013;43(4):813-7.
9. Madani A, Fiore JF Jr, Wang Y, et al. An enhanced recovery pathway reduces duration of stay and complications after open pulmonary lobectomy. Surgery 2015;158(4):899-908.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Lung, Esophageal, Surgery, Surgery : Thoracic Surgery

De.6. Cross Cutting Areas *(check all the areas that apply):*
Safety, Safety : Complications

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.)*
See Appendix

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:** S.15._Detailed_risk_model_specifications_-_0459_Lobectomy_LOS.docx

S.3. For endorsement maintenance, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
To minimize confusion, the measure title was changed from Risk-Adjusted Morbidity: Length of Stay >14 Days after Elective Lobectomy for Lung Cancer to Risk-Adjusted Length of Stay >14 Days after Elective Lobectomy for Lung Cancer.

The denominator time window was changed from 12 months to 36 months. The 36-month time window is necessary to obtain appropriate sample sizes for this measure.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer who had a prolonged length of stay >14 days

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
36 months

S.6. 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, 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.
Prolonged Postoperative Length of Stay (PLOS) is defined as a Yes/No variable indicating postoperative hospital stay of greater than 14 days, using surgery date (SurgDt- STS General Thoracic Surgery Database (GTSD) v 2.2, sequence number 1340) and discharge date (DischDt- STS GTSD v 2.2, sequence number 2190) to calculate PLOS.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Number of patients aged 18 years and older undergoing elective lobectomy for lung cancer

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk : Individuals with multiple chronic conditions, Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)

1. Lung cancer (LungCancer - STS GTS Database, v 2.2, sequence number 830) is marked "yes" and Category of Disease – Primary (CategoryPrim - STS GTS Database, v 2.2, sequence number 1300) is marked as one of the following:
(ICD-9, ICD-10)
Lung cancer, main bronchus, carina (162.2, C34.00)
Lung cancer, upper lobe (162.3, C34.10)
Lung cancer, middle lobe (162.4, C34.2)
Lung cancer, lower lobe (162.5, C34.30)
Lung cancer, location unspecified (162.9, C34.90)

2. Primary procedure is one of the following CPT codes:
Thoracoscopy, surgical; with lobectomy (32663)
Thoracoscopy with removal of two lobes (bilobectomy) (32670)
Removal of lung, single lobe (lobectomy) (32480)
Removal of lung, two lobes (bilobectomy) (32482)
Removal of lung, sleeve lobectomy (32486)

3. Status of Operation (Status - STS GTS Database, v 2.2, sequence number 1420) is marked as "Elective"

4. Gender (Gender -STS GTS Database, v 2.2, sequence number 190) is marked "Male" or "Female," surgery date (SurgDt - sequence number 1340), and discharge date (DischDt – sequence number 2190) are provided

5. Only analyze first operation of hospitalization meeting criteria 1-4.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
None

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
N/A

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The model was developed by multivariate stepwise logistic regression. The details of risk adjustment model development were published in 2008:

Wright CD, Gaissert HA, Grab JD, O´Brien SM, Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.

The following covariates are included in the PLOS model:

Variable	Sequence #	Definition
Age	170	
Gender	190	
Zubrod Score	820	
ASA Class	1470	

Insulin-dependent Diabetes	640, 650	If "Yes" for Diabetes (640) and "Insulin" marked for Diabetes Control (650)
Renal Dysfunction	680	If Creatinine >= 2 or "Dialysis of any type" at time of model development; it is now consistent with RIFLE criteria
Preoperative Therapy	580,600	If "Yes" for Preoperative chemotherapy (580) or "Yes" for Preoperative Thoracic Radiation Therapy (600)
FEV (% Predicted)	790	
Cigarette Smoking (Ever)	730	

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
To maximize use of available data, when encountering records with missing values of the model covariates (with exception of age, gender, and FEV), the missing values are imputed to the most common value of the covariate among the remaining eligible cases. Patient records missing age or gender are excluded. For the continuous variable FEV, missing values are imputed to the median FEV of the remaining eligible cases.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*

IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS General Thoracic Surgery Database (GTSD) Version 2.2; STS GTSD Version 2.3 went live on January 1, 2015.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
2.1_Testing_-_0459_Lobectomy_LOS.4.5.16.docx

PREVIOUS MEASURE TESTING

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0459 **NQF Project:** Clinician Level Perioperative Care

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
STS General Thoracic Surgery Database - Compare results between two consecutive 6-month time intervals during 2009: January 2009 - June 2009 and July 2009 - December 2009

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)*:
To assess temporal reliability of the proposed measure, we estimated hospital rates of prolonged stay for two consecutive 6-month time intervals during 2009 and compared the results. Only 12 months of data were available for the current data version v2.081. Thus, we were unable to consider time intervals longer than 6 months. Only hospitals

with data for both time periods were included. Hospital estimates for each time period were estimated using hierarchical models, as described above. The correlation between hospital estimates in two time periods was assessed graphically and summarized by the Pearson correlation coefficient.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The Pearson correlation between hospital-specific rates of prolonged stay in the 1st vs. 2nd half of 2009 was 0.31.

Site-specific proportions of patients experiencing PLOS>14days (based on hierarchical model)

Correlation between performance in two consecutive time periods

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
STS General Thoracic Surgery Database

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
When data arrive at the data warehouse, they are checked carefully for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report that is generated automatically following each harvest file submission. Upon receipt of the data quality report, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis. If the data inconsistencies are not changed by the participant prior to harvest close, the data warehouse performs consistency edits and/or parent/child edits on the data in order for them to be analyzable. Participants are informed of such edits to their data in the Data Quality Report.

Since 2006, the Iowa Foundation for Medical Care (IFMC) has conducted audits of the STS Adult Cardiac Surgery Database on the behalf of STS. Beginning in the fall of 2010, IFMC will conduct audits of the STS General Thoracic Surgery Database to evaluate the accuracy, consistency and comprehensiveness of data collection which will validate the integrity of the data. 5% of participants will be randomly selected annually. Auditors will validate case inclusion and twenty lobectomy cases will be randomly chosen for review of thirty individual data elements. The auditors will abstract each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates will be calculated for each of the 30 elements as well as an overall agreement rate.

In addition, validity was confirmed and is regularly assessed by an expert panel of thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Task Force on Quality Initiatives and the STS Workforce on National Databases.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results*

demonstrating the need to specify them.)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Please see Risk Adjustment Type section above

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Detailed information regarding the risk adjustment model can be found in the attachment:

Wright CD, Gaissert HA, Grab JD, O´Brien SM , Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Data for this analysis were based on 4269 patients who underwent surgery during 2009 at hospitals participating in the STS General Thoracic Surgery Database and met the inclusion criteria for the measure. Only hospitals submitting operations in both semesters of 2009 were included.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
For each patient, we determined whether the patient stayed longer than 14 days. Hospital-specific probabilities of prolonged stay were estimated in a Bayesian hierarchical model. These model-based estimates were used to control variation due to random statistical fluctuations while estimating true signal variation. Hospital-specific estimates were summarized by percentiles and plotted along with 95% credible intervals to illustrate between-hospital variation.

Results for each participant are presented in two forms: (1) the estimated risk-adjusted rate (RAR); and (2) the estimated standardized incidence ratio (SIR). The RAR is interpreted as the outcome rate that would be observed hypothetically for a participant if the participant performed surgery on each eligible patient in the STS General Thoracic Surgery Database. This hypothetical quantity cannot be observed directly but may be estimated in a statistical model, as described below.

A participant's SIR is defined as the ratio of the participant's RAR divided by the overall STS observed outcome rate.
SIR of participant = RAR of participant / overall STS observed rate.

An SIR value greater than 1.0 implies that the participant's risk-adjusted outcome rate is higher than the overall STS observed rate. Conversely, an SIR value less than 1.0 implies that the participant's risk-adjusted outcome rate is lower than the overall STS observed rate.

To account for uncertainty in the estimation of RAR and SIR, the estimates of these quantities are accompanied by 95% credible intervals (CI). The 95% CI indicates the range of RAR and SIR values that are plausible in light of the observed data. If the 95% CI for a participant's SIR includes the null value 1.0, then we cannot reliably distinguish this participant's performance from the STS average - either the participant's performance was close to average or else the participant's sample size was too small to make a reliable determination.

Statistical Model
Random effects logistic regression models were used to compare each participant's event rate in a manner that adjusts for case mix and accounts for uncertainty due to small sample sizes. Random effects models use data from all database participants when estimating the event rate of a single participant, thereby borrowing strength to obtain a more reliable estimate. Each outcome was adjusted for its own set of patient
factors (listed below) and included a separate random effect parameter for each participant in the analysis. The model has the form:

where pji denotes the probability of prolonged stay for the i-th patient at participant j; ej denotes a (random effect) intercept parameter for participant j; and xjiq denotes the value of q-th covariate for the i-th patient at the j-th participant. The terms xjiq represent quantitative risk factors such as age and FEV; and binary indicator variables (e.g. 1=male, 0=female).

Parameters of the random effects logistic model were estimated in a Bayesian framework using WinBUGS software. Unlike conventional statistical methods, the results of Bayesian analyses are expressed in terms of probabilities. For example, we might be 95% sure that a participant has a better-than-average risk-adjusted rate. The Bayesian 95% credible interval (CI) has the following interpretation. In light of the observed data, it is 95% likely that the true value of the parameter of interest lies in the indicated interval.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
Prolonged length of stay occurred in 5.1% of eligible patients. Hospital-specific estimates ranged from 2.6% to 14.4%.

Distribution of hospital-specific estimated probabilities of prolonged length of stay > 14 days

Min	25th percentile	Median	75th percentile	Maximum
2.6%	4.1%	4.9%%	6.2%%	14.4%%

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*: **2b6.3 Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
2c. Disparities in Care: H☐ M☐ L☐ I☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)
2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: **2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
2.1-2.3 Supplemental Testing Methodology Information:
Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? (*Reliability and Validity must be rated moderate or high*) Yes☐ No☐ **Provide rationale based on specific subcriteria:**
If the Committee votes No, STOP

UPDATED MEASURE TESTING

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0459
Measure Title: Risk-Adjusted Length of Stay >14 Days after Elective Lobectomy for Lung Cancer
Date of Submission: 3/14/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.

- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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*).

STS General Thoracic Surgery Database (GTSD) Version 2.2

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

July 1, 2012 – June 30, 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.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: 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*)

The calculation of this measure using the 36 months from July 1, 2012 to June 30, 2015 included 23,174 operations from 244 STS General Thoracic Surgery Database participants. Description of the distribution of participant size (patient volume) overall and by geographic region is below

	Total	**Midwest**	**Northeast**	**South**	**West**
# of participants	244	49	73	82	40
# of patients	23174	5020	7756	7402	2996

Participant size (volume)

Mean	95	102.4	106.2	90.3	74.9
SD	94.4	104.9	118.6	75.7	57.8
IQR	107	101	128	97.8	76
Minimum	1	3	1	1	5
10% percentile	10	10	11	15	10
20% percentile	22	24	21	27	22
30% percentile	36	33	35	42	34
40% percentile	49	47	46	60	48
Median	67	67	64	71	56
60% percentile	88	86	90	89	77
70% percentile	114	112	141	113	100
80% percentile	152	174	163	149	122
90% percentile	206	253	256	196	167
Maximum	626	411	626	399	208

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*)

Includes 23,174 eligible patients. Patient characteristics are below.

Age (years), mean (SD)	67.2 (9.7)
Male	45.1%
Zubrod score	
0	45.1%
1	51.0%
2	3.1%
3	0.7%
4-5	0.1%
ASA Class	
I	0.2%
II	14.9%
III	75.8%
IV-V	9.2%
Diabetes mellitus	4.2%
Renal dysfunction	0.3%
Induction therapy	10.2%
Smoking (ever)	84.8%
$\%FEV_1$, mean (SD)	83.5 (19.7)

Year of surgery	
2012	15.5%
2013	33.9%
2014	34.7%
2015	15.8%

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.

No data differences are present.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

No SDS variables were collected and utilized.

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 is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. Accordingly, reliability was defined as the square of the Pearson correlation coefficient (ρ^2) between the set of participant-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$ and the corresponding unknown true values $\theta_1, \ldots, \theta_N$ (N=number of participants), that is:

$$\rho^2 = \frac{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)\left(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h\right)}{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)^2 \sum_{j=1}^{N}\left(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h\right)^2}.$$

The quantity ρ^2 was estimated by its posterior mean estimated from a MCMC chain of length 2000, namely, $\hat{\rho}^2 = \frac{1}{2000}\sum_{l=1}^{2000}\rho_{(l)}^2$ where

$$\rho^2_{(l)} = \frac{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)\left(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)}\right)}{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)^2 \sum_{j=1}^{N}\left(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)}\right)^2}$$

with $\theta_j^{(l)}$ denoting the value of θ_j on the l-th MCMC sample and $\hat{\theta}_j = \frac{1}{2000}\sum_{l=1}^{2000}\theta_j^{(l)}$ denoting the estimated posterior mean of θ_j. A 95% credible interval for ρ^2 was obtained by calculating the 50-th smallest and 50-th largest values of $\rho^2_{(l)}$ across the 2000 MCMC samples ($l = 1, \ldots, 2000$).

Kozower BD, O'Brien SM, Kosinski AS, Magee MJ, Dokholyan R, Jacobs JP, Shahian DM, Wright CD, Fernandez FG. The Society of Thoracic Surgeons Composite Score for Rating Program Performance for Lobectomy for Lung Cancer. Ann Thorac Surg. 2016 Jan 16. pii: S0003-4975(15)01753-1. doi: 10.1016/j.athoracsur.2015.10.081. [Epub ahead of print]. PMID: 26785936.

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*)

Based on the 244 participants, the reliability (proportion of signal variation) is 37.6% with a 95% credible interval [CrI] (26.1%, 48.9%). Reliability increases when considering participants with a particular minimum number of procedures per year as displayed below.

	All participants	**≥10 procedures per year**	**≥20 procedures per year**	**≥30 procedures per year**	**≥40 procedures per year**
Number of participants	244	184	132	95	72
Reliability	37.6%	44.5%	49.5%	56.1%	63.8%
95% CrI for reliability	(26.1%, 48.9%)	(32.3%, 55.5%)	(36.2%, 61.5%)	(41.3%, 68.6%)	(48.6%, 76.4%)

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 improves with volume and it is comparable to reliability for other measures for STS GTSD.

2b2. VALIDITY TESTING

2b2.1. 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 <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level 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)*

Critical data elements

When data arrive at the data warehouse, they are checked for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report (DQR) that is generated automatically following each harvest file submission. Upon receipt of the DQR, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis.

In addition, participating sites are randomly selected for participation in the STS General Thoracic Surgery Database (GTSD) Audit, which is designed to evaluate the accuracy, consistency, and comprehensiveness of data collection and ultimately validate the integrity of the data contained in the database. Telligen, formerly the Iowa Foundation for Medical Care, has conducted audits on behalf of STS since 2006. In 2015, ten percent of randomly selected STS GTSD participants (N = 25, an increase from 24 in 2014 and 18 in 2013) were audited. The audit process involves re-abstraction of data for 20 cases records (at least 15 lobectomy and up to 5 esophagectomy) and comparison of 40 STS GTSD V2.2 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each variable, each variable category and overall. In 2015, the overall aggregate agreement rate was 97.02%, demonstrating that the data contained in the STS GTSD are both comprehensive and highly accurate.

Data Analysis

Aggregate agreement rates were computed for all facilities by calculation of the sum of all facilities' numerators divided by the sum of all facilities' denominators, for each individual variable, each variable category and overall.

Chi-square statistics were calculated to identify any possible relationships between the data collection process variables and agreement rates. Tests where the chi-square statistic had a probability of less than 5% ($p<0.05$) were considered to show statistically significant differences in agreement rate between the levels of the process measure.

Agreement Rate Results

Database validity was evaluated by re-abstraction of defined variables from the medical records and comparison to submitted data. Agreement rates were calculated at the individual variable level, category level and overall. Aggregate agreement rates are presented in the table below.

CATEGORY	FIELD_NAME	NUM	DEN	Agreement Rate
PRE-OPERATIVE EVALUATION	OVERALL_ALL FIELDS	6213	6452	96.30%
Pre-Operative Evaluation	Admission Date	496	500	99.20%
Pre-Operative Evaluation	Prior Cardiothoracic Surgery	489	500	97.80%
Pre-Operative Evaluation	Pre-Op Chemo-Current Malignancy	494	500	98.80%
Pre-Operative Evaluation	Pre-Op Thoracic Radiation Therapy	495	500	99.00%
Pre-Operative Evaluation	Diabetes	409	415	98.55%
Pre-Operative Evaluation	Diabetes Control	69	82	84.15%
Pre-Operative Evaluation	Cigarette Smoking	475	500	95.00%
Pre-Operative Evaluation	Pulmonary Function Tests Performed	394	415	94.94%
Pre-Operative Evaluation	FEV1 Predicted	349	377	92.57%
Pre-Operative Evaluation	Zubrod Score	467	499	93.59%

CATEGORY	FIELD_NAME	NUM	DEN	Agreement Rate
Pre-Operative Evaluation	Lung Cancer	408	409	99.76%
Pre-Operative Evaluation	Clinical Staging Method-Lung-EBUS	355	366	96.99%
Pre-Operative Evaluation	Clinical Staging Method-Lung-PET or PET/CT	337	366	92.08%
Pre-Operative Evaluation	Lung Cancer Tumor Size	338	364	92.86%
Pre-Operative Evaluation	Lung Cancer Nodes	354	361	98.06%
Pre-Operative Evaluation	Esophageal Cancer	85	85	100.0%
Pre-Operative Evaluation	Clinical Staging Method-Esophageal-EUS	62	71	87.32%
Pre-Operative Evaluation	Esophageal Cancer Tumor	70	71	98.59%
Pre-Operative Evaluation	Esophageal Cancer Nodes	67	71	94.37%
Diagnosis And Procedures	OVERALL_ALL FIELDS	4663	4809	96.96%
Diagnosis And Procedures	Category of Disease-Primary	493	500	98.60%
Diagnosis And Procedures	Date of Surgery	500	500	100.0%
Diagnosis And Procedures	Procedure Start Time	481	500	96.20%
Diagnosis And Procedures	Procedure End Time	463	500	92.60%
Diagnosis And Procedures	ASA Classification	481	500	96.20%
Diagnosis And Procedures	Procedure	493	500	98.60%
Diagnosis And Procedures	Patient Disposition	481	500	96.20%
Diagnosis And Procedures	Pathological Staging-Lung Cancer-T	355	359	98.89%
Diagnosis And Procedures	Pathological Staging-Lung Cancer-N	357	359	99.44%
Diagnosis And Procedures	Lung Cancer - Number of Nodes	339	359	94.43%
Diagnosis And Procedures	Pathological Staging-Esophageal Cancer-T	86	90	95.56%
Diagnosis And Procedures	Pathological Staging-Esophageal Cancer-N	69	71	97.18%
Diagnosis And Procedures	Esophageal Cancer-Number of Nodes	65	71	91.55%
POST-OPERATIVE EVENTS	OVERALL_ALL FIELDS	1586	1598	99.25%
Post-Operative Events	Return to OR	496	500	99.20%
Post-Operative Events	Pneumonia	494	500	98.80%
Post-Operative Events	Initial Vent Support >48 Hours	498	500	99.60%
Post-Operative Events	Gastric Outlet	98	98	100.0%
DISCHARGE	OVERALL_ALL FIELDS	1950	1995	97.74%
Discharge	Discharge Date	500	500	100.0%
Discharge	Discharge Status	494	500	98.80%
Discharge	Readmission within 30 Days of Discharge	486	495	98.18%
Discharge	Status at 30 Days	470	500	94.00%
	OVERALL_ALL FIELDS	14412	14854	97.02%

There were 14,854 total variables abstracted and of those 14,412 variables matched, resulting in an overall agreement rate of 97.02%.

Process Variable Correlation Tables

The relationships between process variables and overall agreement rates were examined and included:

- Facility data collection performed from electronic medical records or a combination of paper and electronic medical records and overall agreement rate
- Facility data collection method (concurrent/retrospective/both) and overall agreement rate
- Data collection performed by a single abstractor or multiple staff and overall agreement rate
- Attendance at the annual data managers' meeting, STS Advances in Quality and Outcomes (AQO) Conference, and overall agreement rate
- Agreed upon abstraction location for data elements documented in multiple locations and overall agreement rate

Relationship between Data Collection Source & Agreement Rate
Facilities using an electronic health record (EHR) for data collection had higher agreement rates, 97.36%, than those facilities using both paper medical records and an EHR, 96.31%. There were no facilities that used paper medical records alone (p<0.0004).

Relationship between Data Collection Method & Agreement Rate
Facilities collecting data retrospectively have higher agreement rates, 97.55%, than those facilities collecting data concurrently, 96.18%, or both, 96.38% (p=<0.0001).

Relationship between Data Collection Performed By & Agreement Rate
Facilities with a single individual performing data abstraction have higher agreement rates, 98.02%, than those facilities that have multiple individuals performing data abstraction, 96.24% (p<0.0001).

Relationship between Attendance at AQO Conference & Agreement Rate
Facilities having staff attend the annual AQO Conference have higher agreement rates, 97.25%, than those that do not have staff attend, 96.11% (p<0.0012).

Relationship between Have an Agreed Upon Location & Agreement Rate
Facilities that utilize an agreed upon location for data elements recorded in multiple locations have higher agreement rates, 97.31%, than facilities that do not utilize an agreed upon location, 93.61% (p<0.0001).

In addition, validity is regularly assessed by an expert panel of general thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Quality Measurement Task Force, and the STS Task Force on Quality Initiatives, all of which report to the STS Workforce on National Databases.

Performance measure score
The measure is regarded as useful and valid by its intended users. The measure was developed with a panel of surgeon experts and statisticians. Differences in the measure across participants are clinically meaningful.

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

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

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

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

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

measure scores)

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

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 10 risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Covariates were selected a-priori based on clinical relevance, literature review, and empirical analysis by a panel of physicians and statisticians. The details of model development were published in 2008 (Wright et al).

Risk of mortality and other short-term clinical outcomes is mostly influenced by clinical factors present on admission, such as age, comorbidities, and pulmonary function. By convention, given the plausible causal pathways leading to these outcomes, risk models used for mortality profiling have generally excluded clinical patient factors or local environmental factors, as their inclusion might theoretically adjust out important inequities in care.

Prolonged length of stay after lobectomy could theoretically be impacted by sociodemographic factors. There is very little written in the literature about sociodemographic factors as they relate to length of stay after lobectomy. Giambrone et al used state inpatient databases from Florida, California, and New York to study length of stay in patients who had uncomplicated pulmonary lobectomies. There was significant variability in LOS even for uncomplicated operations. Typical clinical factors were shown to prolong LOS but so did Medicaid vs. private insurance; with Medicaid being an independent predictor of longer length of stay. Additionally, Stitzenberg et al examined the likelihood of undergoing thoracoscopic lobectomy in the mid-Atlantic states between 2007 and 2008. This study demonstrated that patients in rural areas, in low-volume hospitals, who had Medicaid, or with lower median incomes underwent VATS lobectomy less often. VATS lobectomy was associated with shorter hospital stay in this study. Given the lack of consistent, compelling evidence regarding SDS factors and LOS, they are not included in the current measure being submitted.

Wright CD, Gaissert HA, Grab JD, O'Brien SM , Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.

Giambrone GP, Smith MC, Wu X et al. Variability in length of stay after uncomplicated pulmonary lobectomy: is length of stay a quality metric or a patient metric? Eur J Cardiothorac Surg 2016 April; 49(4):e65-71.

Stitzenberg, KB, Shah PC, Snyder JA, Scott WJ Disparities in access to video-assisted thoracic surgical lobectomy for the treatment of early-stage lung cancer. J Laparoendosc Adv Surg Tech A 2012 Oct; 22(8):753-7.

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

Prolonged Length of Stay after Lobectomy for Lung Cancer Model		
Variable	**OR (95% CI)**	**p**
Age (per 10 yr increase)	1.213 (1.127, 1.305)	<0.001
Male	1.587 (1.385, 1.818)	<0.001
Zubrod score	1.355 (1.224, 1.500)	<0.001
ASA score	1.542 (1.341, 1.772)	<0.001
Diabetes mellitus	1.031 (0.770, 1.381)	0.84
Renal dysfunction	1.621 (0.731, 3.595)	0.26
Induction therapy	1.241 (1.012, 1.522)	0.043
Smoking (ever)	1.432 (1.121, 1.829)	0.003
%FEV_1 (per 10% increase)	0.853 (0.825, 0.883)	<0.001
Year of surgery (per 1 yr increase)	0.938 (0.874, 1.005)	0.071

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)*

No sociodemographic (SDS) variables were collected and utilized.

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

The model was assessed for discrimination by means of C-index and for goodness-of-fit through Hosmer-Lemeshow statistic.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
C-statistic is 0.672.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.94 (Chi-Square=2.89, df=8)

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not provided

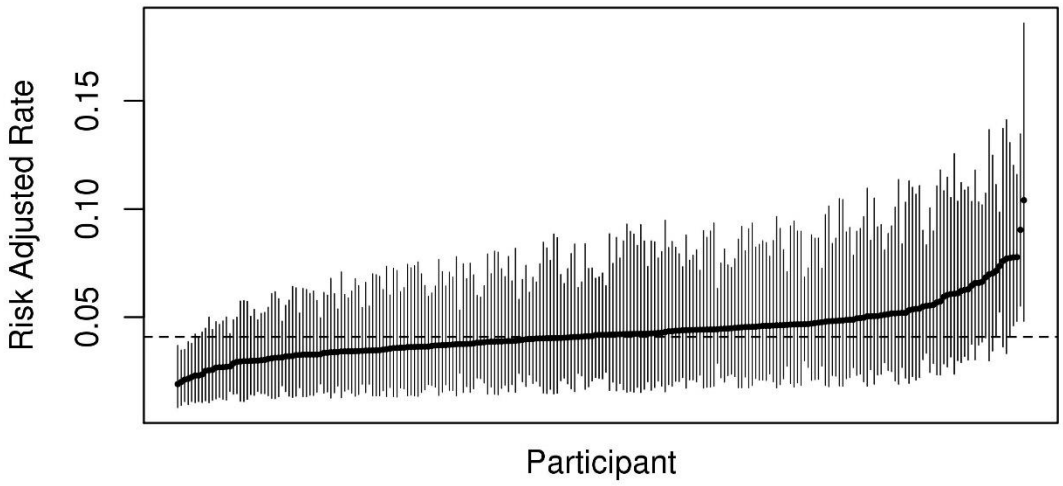
2b4.9. Results of Risk Stratification Analysis:
N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
The results demonstrated that the risk model is well calibrated and has good discrimination power. It is suitable for controlling differences in case-mix between participants.

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

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
Participant-specific risk adjusted rates (RAR) for the prolonged length of stay after lobectomy were estimated within a Bayesian hierarchical logistic regression model. Participant-specific RARs were plotted along with corresponding 95% credible intervals to illustrate the between-participant variation.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
Better than expected and worse than expected participants are distinguishable because they have the 95% credible intervals below or above the STS average (dashed line) as seen on the plot below.



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

The identified differences in performance are clinically meaningful and allow differentiation between participants based on estimates and 95% credible intervals.

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

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

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

2b6.3. 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*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

The quality of data in STS General Thoracic Surgery Database has been improving. We managed the missing data with imputation. Missing %FEV1 values were imputed utilizing median of the observed %FEV1 values. For binary risk factors, missing values were considered as indicating absence of the risk factor.

2b7.2. 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*)

Variables with missing data were: renal dysfunction (4.49%), induction therapy (2.41%), %FEV1 (4.28%), smoking status (0.02%). Other covariates had no missing data.

2b7.3. 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*)

The missing data approach was compared, as a sensitivity analysis, to multiple imputations and the results were insensitive to the approach. Thus, the simpler approach was taken as explained in the reference below.

Wright CD, Gaissert HA, Grab JD, O'Brien SM , Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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.**
The STS General Thoracic Surgery Database (GTSD) has more than 270 participants, and local availability of data elements in electronic format varies across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS GTSD in an electronic format following a standard set of data specifications. STS GTSD participating institutions utilize data entry software products that are approved for the purposes of collecting and submitting STS GTSD data elements.

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.**
No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The data elements included in this measure have been standard in the STS General Thoracic Surgery Database for at least 3 years and some of them have been part of the database for more than 15 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care. Every 3 years, the STS GTSD undergoes a specification upgrade (i.e., a process including surgeon leadership, staff, database managers, and programmers to review and update data fields and their respective definitions to ensure they reflect current practice).

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)*.
Data Collection:
There are no direct costs to collect data for this measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon leaders' time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS General Thoracic Surgery Database participants (single surgeon or a group of surgeons) pay annual fees of $550 per surgeon for STS members and $700 per surgeon for non-STS members. In addition, there is a cost associated with purchasing data collection software which varies across vendors. STS GTSD participants have a separate agreement with their vendor, a process in which STS is not involved.

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.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) STS General Thoracic Surgery Database – 273 Participants http://www.sts.org/national-database/database-managers/general-thoracic-surgery-databas Quality Improvement (Internal to the specific organization) STS General Thoracic Surgery Database – 273 Participants http://www.sts.org/national-database/database-managers/general-thoracic-surgery-databas

4a.1. **For each CURRENT use, checked above, provide:**

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

Please see above

4a.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?*)

In 2017, STS is planning to launch the general thoracic surgery component of STS Public Reporting Online. This is currently in the planning stage. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online

The STS National Database is a Qualified Clinical Data Registry (QCDR). Prolonged length of stay is one of many measures that STS reports to CMS on behalf of consenting STS Adult Cardiac Surgery Database surgeons. For the STS General Thoracic Surgery Database (GTSD), physician quality reporting is in the planning stage. STS GTSD leaders and STS staff are reviewing general thoracic measures for inclusion in physician quality reporting. STS intends to include general thoracic measures in its 2017 QCDR self-nomination.

4a.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.*)

Please see 4a.2.

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Please see 1b.2 and 1b.4

The performance of each participant is calculated twice a year using an updated 3-year time window. The performance is compared to the average STS performance within the considered 3 year time window. These results are shared with all participants through a semiannual feedback report.

	Total	Midwest	Northeast	South	West
# of participants	244 (100%)	49 (20.1%)	73 (29.9%)	82 (33.6%)	40 (16.4%)

# of patients	23174 (100%)	5020 (21.7%)	7756 (33.5%)	7402 (31.9%)	2996 (12.9%)

4b.2. **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.**
N/A

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences.

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

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 completely harmonized?

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

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:** Appendix_-_0459_Lobectomy_LOS.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

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.
• David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
• Gaetano Paone, MD – Chair, Task Force on Quality Initiatives (TQI); surgeon leader/clinical expert in adult cardiac surgery
• Benjamin Kozower, MD – Chair, General Thoracic Surgery Database Task Force; surgeon leader/clinical expert in general thoracic surgery
• William Burfeind, MD – Surgeon leader/clinical expert in general thoracic surgery
• Robert Welsh, MD – Surgeon leader/clinical expert in general thoracic surgery
• Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
• Felix Fernandez, MD – Surgeon leader/clinical expert in general thoracic surgery
• Cameron Wright, MD – Surgeon leader/clinical expert in general thoracic surgery
• Max He, MS – Statistician
• Sean O´Brien, PhD – Statistician
• Andrzej Kosinski, PhD – Statistician
• Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
• Donna McDonald, MPH, RN – Staff, Senior Manager of the STS National Database and Patient Safety

Members of the STS TQI and the GTSDTF provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

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:** 03, 2016
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?** 01, 2017

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** None

S.15. Detailed risk model specifications

Table 1. Predictors of Prolonged Length of Stay After Lobectomy for Lung Cancer

Variable	Estimated OR	95% CI	Bayesian Probability[a]
Age, 10-year increment	1.30	1.15–1.47	<0.001
Zubrod score	1.51	1.27–1.77	<0.001
Male gender	1.45	1.13–1.81	0.002
ASA score	1.54	1.22–1.88	<0.001
Diabetes mellitus	1.71	1.10–2.80	0.037
Renal dysfunction	1.79	1.14–2.60	0.004
Induction therapy	1.65	1.19–2.21	0.001
$\%FEV_1$ (10% change)	0.88	0.83–0.94	<0.001
Smoking (ever)	1.33	0.88–1.96	0.095
Year of surgery	0.99	0.90–1.10	0.555

[a] The Bayesian probability is the probability that the true association with the outcome is on the opposite side of the null hypothesis value from the estimated value; for example, the probability that the true odds ratio for diabetes is < 1.0 given the observed value of 1.71 is 0.037.

ASA = American Society of Anesthesiology; CI = confidence interval; $\%FEV_1$ = the percentage predicted forced expiratory volume in 1 second; OR = odds ratio.

The model adjusted for the 10 patient factors listed above and included a separate random effect parameter for each of the participants in the analysis. The model has the form:

$$\log(p_{ji}/[1-p_{ji}]) = \beta_0 + \beta_1 x_{ji1} + \cdots + \beta_q x_{jiq} + e_j$$

where p_{ji} denotes the probability of prolonged stay for the i-th patient at participant j; e_j denotes a (random effect) intercept parameter for participant j; and x_{jiq} denotes the value of q-th covariate for the i-th patient at the j-th participant. The terms x_{jiq} represent quantitative risk factors such as age and $\%FEV_1$ and binary indicator variables (eg, 1 = male, 0 = female).

STS GTSD data definitions are provided in S.1. Detailed risk model specifications are provided in the attached manuscript:

Wright CD, Gaissert HA, Grab JD, O'Brien SM, Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg. 2008 Jun;85(6):1857-65.

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 #: 0460
Measure Title: Risk-Adjusted Morbidity and Mortality for Esophagectomy for Cancer
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: Percentage of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer who developed any of the following postoperative conditions: bleeding requiring reoperation, anastomosis leak requiring medical or surgical treatment, reintubation, ventilation >48 hours, pneumonia, or discharge mortality
Developer Rationale: It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Measuring risk adjusted morbidity and mortality of patients undergoing esophagectomy for cancer provides surgeons and institutions the opportunity to evaluate outcomes and subsequently design quality improvement initiatives to address identified deficits. Utilization of the insight gained should promote improved patient outcome.

Numerator Statement: Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer who developed any of the following postoperative conditions: bleeding requiring reoperation, anastomosis leak requiring medical or surgical treatment, reintubation, ventilation >48 hours, pneumonia, or discharge mortality.
Denominator Statement: Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer
Denominator Exclusions: None

Measure Type: Outcome
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Jul 31, 2008

Maintenance of Endorsement -- Preliminary Analysis

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 outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer stated that esophageal cancer is an aggressive disease with a generally poor prognosis. The incidence of esophageal adenocarcinoma is increasing faster than any other malignancy in the United States. In 2015, there were an estimated 16,980 people diagnosed with esophageal cancer.
- Enhanced recovery pathways and fast-track protocols, have been shown to reduce major morbidity and length of

stay without increasing mortality or readmissions. Knowing their rate of risk adjusted morbidity and mortality after esophagectomy gives thoracic programs the opportunity to design quality improvement initiatives around deficiencies.

Guidance from the Evidence Algorithm: Health outcome measure→The relationship between the outcome and at least one process is identified and supported by the stated rationale→Pass

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

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 the following performance data from the STS General Thoracic Surgery Database (GTSD) for patients that underwent elective esophagectomy for primary esophageal cancer between July 1, 2012 and June 30, 2015.

	Total	Midwest	Northeast	South	West
# of participants	169	35	53	54	27
# of patients	4557	1325	1499	1176	557

Participant-specific risk adjusted rates (RAR)

Mean	29.3	30.2	28.9	29.5	28.6
SD	7.3	8.8	7.2	6.5	7.1
IQR	8.3	9.8	7.4	7.5	7.4
Minimum	13.5	15.3	13.5	14.5	16.2
10% percentile	20.7	20.6	20.6	22.5	21.2
20% percentile	24.3	23.1	24.4	24.8	23.3
30% percentile	25.8	26.3	26.0	26.1	24.5
40% percentile	27.2	28.1	27.0	27.2	25.7
Median	27.9	28.6	27.7	27.9	27.7
60% percentile	30.3	31.8	30.0	30.2	30.3
70% percentile	32.3	32.9	32.2	32.3	31.3
80% percentile	34.4	34.8	34.5	34.5	33.4
90% percentile	38.5	42.6	38.1	39.1	34.3
Maximum	51.7	51.7	45.7	44.3	47.2

- The developer stated that the endpoint of mortality or major morbidity occurred in **27.3%** (1,243/4,557) of eligible patients from July 1, 2012 through June 30, 2015.
- The developer did not provide individual rates for the postoperative conditions that define morbidity (bleeding requiring reoperation, anastomotic leak requiring medical or surgical treatment, reintubation, initial ventilation >48 hours, or pneumonia) and mortality.
- For endorsement maintenance, NQF asks for performance scores on the measure as specified*, current and over time*. The developer did not provide performance data on the measure from 2008 (when first endorsed) through 2012.

Disparities:

- The developer did not provide data on disparities from the measure as specified – this is encouraged for endorsement maintenance.

Questions for the Committee:

- *Without performance data prior to 2012, is it possible to determine if there is a quality problem and opportunity for improvement in care for patients that underwent elective esophagectomy for primary esophageal cancer that warrants a national performance measure?*
- *Are you aware of any disparities data that exists 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 to Support Measure Focus

Comments:

This is an outcome measure. It does not specifically link to any one specific structure, process, intervention, or service. Although the developers of the measure indicate that the use of ERPs may improve on this number, there is no specific evidence given that a performance gap exists in the use of ERPs

**As enhanced recovery pathways or fast-track protocols, appear to reduce major morbidity and length of stay without increasing mortality or readmissions, there does appear to be identified strategies that may impact outcomes. **

1b. Performance Gap

Comments:

**Performance data is given, and given the relatively high morbidity and mortality of patients undergoing esophagectomy, one would assume that there is room for improvement. However, the measure does not in any way measure any specific process that might lead to such an improvement, and essentially relies on the Hawthorne effect to get there. **

**Disparities by race are in the risk model. SES disparities are not in the model. This indicates that race disparities will be under-reported in this circumstance since they are risk adjusted away. SES variables are not collected so there is little opportunity to assess this. **

**The Participant-specific risk adjusted rates, which were provided, appear to demonstrate fairly substantial variation in outcomes. This range from the 2012-2015 analysis appears similar to aggregate data reported in the 2009 article by Wright et al. It appears that a performance gap persists. **

**Definitive data on disparities among population subgroups isn't available, although the risk adjustment and an analysis provided by the developer suggest that age, gender, and race are relevant. **

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Electronic Clinical Data : Registry

Specifications:

- This is a clinician-level measure.
- The measure is risk adjusted.
- The numerator of this measure is: Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer who developed any of the following postoperative conditions: bleeding requiring reoperation, anastomosis leak requiring medical or surgical treatment, reintubation, ventilation >48 hours, pneumonia, or discharge mortality.
- The denominator is: Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer.
- The denominator time window was changed from 12 months to 36 months. The 36-month time window is

necessary to obtain appropriate sample sizes for this measure.
- There are no exclusions.
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The developer refers to numerator and denominator sections for detailed information about the calculation algorithm.
- The developer addresses how missing data are handled.
- STS General Thoracic Surgery Database (GTSD) is the registry identified as the specific data source for this measure.
- Collection instrument available at measure specific web page.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure consistently implemented?*

2a2. Reliability Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- In the prior submission, the developer assessed test-retest reliability by comparing the results of estimated hospital rates mortality or major morbidity between two consecutive 6-month time intervals during 2009. The Pearson correlation between hospital-specific rates of mortality or majory morbidity in the 1st vs. 2nd half of 2009 was 0.50.
- The previous committee noted that they had concerns that combining morbidity and mortality is not appropriate (i.e., do stakeholders view and value prolonged intubation the same as death?) and may be problematic, although death is one type of morbidity. The committee evaluated how the morbidity and mortality measures perfomed together and separately.

Describe any updates to testing:

- Reliability of the measure score was not presented in prior submission(s), reliability testing of the measure score has been conducted this review.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 4,557 operations from 169 STS General Thoracic Surgery Database participants from July 1, 2012 to June 30, 2015 (36 months).
- The developers conducted a signal-to-noise analysis using the Pearson correlation coefficient to test the measure score reliability; this is an appropriate method. The signal 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.
    - Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Guide from Evans (1996) suggests for the absolute value of r: 0.00-0.19 as “very weak”, 0.20-.039 “weak”,0.40-0.59 “moderate”, 0.60-0.79 “strong”, and 0.80-1.0 “very strong”. [Evans, J.D. (1996) Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove.]

Results of reliability testing:

- The developers provided the reliability results below and noted that the reliability of the measure score increased as the volume of minimum procedures per year for participants increased .

	All participants	≥5 procedures per year	≥10 procedures per year	≥15 procedures per year	≥20 procedures per year
Number of participants	169	75	50	29	20
Reliability	44.4%	67.9%	71.1%	72.5%	80.6%
95% CrI for reliability	(34%, 53.8%)	(55.7%, 77.9%)	(57.8%, 81.8%)	(56.9%, 84.9%)	(65.5%, 91.1%)

Guidance from the Reliability Algorithm: Precise specifications (Box 1)→empiric reliability testing (Box 2) → performance measure score (Box 4) → signal-to-noise analysis used to calculate reliability rates (Box 5) → moderate certainty/confidence that performance measure scores are reliable (Box6b) →Moderate

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is it likely this measure is consistently implemented?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- The developers previously stated that beginning in the fall of 2010 they would conduct patient-level data element validity testing . Twenty randomly selected lobectomy cases previously submitted to the STS data warehouse would be chosen for review of 30 individual data elements. The developer also stated that validity was confirmed by an expert panel of thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Task Force on Quality Initiatives and the STS Workforce on National Databases. No testing results were provided.

Describe any updates to validity testing: see empirical validity testing below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**
☐ **Empirical validity testing of the measure score**

Validity testing method:

- The dataset used for data element validity testing included 10% of randomly selected STS GTSD participants from 2013 to 2015: N = 25 (2015); N = 24 (2014); N = 18 (2013). Twenty cases (at least 15 lobectomy and up to 5 esophagectomy) that were previously submitted to the STS data warehouse were re-abstracted and compared to the 'gold standard'; this is an appropriate method. Agreement rates were calculated for 40 STS GTSD V2.2 individual data elements.
- The developer provided data on the relationship between process variables and agreement rate, though, this does not meet NQF validity testing requirements.
- The developer also stated that the measure is regarded as useful and valid by its intended users and differences in the measure across participants are clinically meaningful; however, the information provided is not sufficient to meet NQF's requirement for face validity.

Validity testing results:

- The developer stated that in 2015, there were 14,854 total variables abstracted and of those 14,412 variables matched. Individual data elements were included in the following categories: pre-operative evaluation, diagnosis and procedures, post-operative events, and discharge.
- Agreement rates for the individual data elements ranged from 84.15% (diabetes control) to 100.0% (esophageal cancer, date of surgery, gastric outlet, and discharge date).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer did not identify any exclusions

Questions for the Committee:

- *Should exclusions be identified for this measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary:

Description of the model

- The measure is risk adjusted using a logistic regression model. The developer stated that covariates were selected based on clinical relevance, literature review, and empirical analysis conducted by a panel of physicians and statisticians.
- The developer did not provide the details of the statistical methods and criteria used to select patient factors in the risk model, though they state that the details of the model development were published in 2009 (Wright et al). The final model includes 16 variables.
- The developer used the C-index and Hosmer-Lemeshow Goodness-of-Fit statistical methods to assess model discrimination calibration but did not provide a detailed description of the analyses.
    - The C-index or c-statistic, reflects how accurately a statistical model is able to distinguish between a patient with an outcome and a patient without an outcome. C-statistic values can range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of patients with and without the outcome of interest and a value of 1.0 indicates that the model perfectly identifies those with and without the outcome of interest. Generally, a c-statistic of at least 0.70 is considered acceptable.
    - The Hosmer-Lemeshow test is used to determine the goodness of fit of the logistic regression model.

Performance of the model

- The developer reported the following statistical results:
  - C-statistic: **0.614**
  - Hosmer and Lemeshow Goodness-of-Fit Test: p-value=**0.62** (Chi-Square=**6.27**, df=**8**)
  - The developer did not provide risk decile plots.

SDS Conceptual Description

- The developer performed a literature search to help inform their conceptualization of the pathways by which SDS factors affect outcomes after majory cancer surgery. The developer stated that there is mixed evidence in the literature linking SDS to poorer outcomes after esophagetcomy, in particular.
- The developer provided the following three SDS factors and esophagetcomy outcomes from the literature:
  - Patients in the lowest quintile of SES have significantly increased rates of failure to rescue from a major complication after undergoing esophagectomy [and other major surgeries]. However, when controlling for patient and hospital factors, SES for esophagectomy was no longer a predictor of failure to rescue in this cohort.
  - Sepsis/sepsis-associated mortality among major cancer surgeries, including esophagectomy and insurance status - the odds of sepsis were highest among esophagectomy patients and those with non-private insurance.
- The developer also provided a study that looked at deprivation scores and operative mortality in the UK.
- The developer stated that given the mixed evidence supporting SDS as a risk factor, it was **not** included in the model.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care?*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Participant-specific risk-adjusted rates (RAR) of the endpoint were estimated within a Bayesian hierarchical logistic regression model. Participant-specific RARs were plotted along with corresponding 95% credible intervals to illustrate the between-participant variation.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- N/A

2b7. Missing Data

- The developer managed the missing data with imputation. For binary risk factors, missing values were considered as indicating absence of the risk factor.
- Variables with missing data were: congestive heart failure (4.59%), coronary artery disease (4.43%), peripheral vascular disease (4.63%), insulin diabetes (5.24%), hypertension (4.32%), steroid use (7.83%), BMI = 4.63%, induction therapy (4.08%), renal dysfunction (7.22%). Other covariates had no missing data.
- The missing data approach was compared, as a sensitivity analysis, to multiple imputations and the results were insensitive to the approach.

Guidance from Validity Algorithm: Precise specifications (Box 1)→ potential threats to validity assessed (Box 2)→

empirical validity testing (Box 3)→validity testing conducted with patient-level data elements (Box 10) →Data element validity compared to gold standard for 40 individual data elements (Box 11)→High/Moderate certainty or confidence that data used in the measure are valid (Box 12a) →Moderate (Moderate is highest eligible rating)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

This should be quite reliable given adequate sample size. The definitions for morbidity and mortality are well constructed and acceptable. Unfortunately many providers do very low volumes leading to issues for widespread implementation

one score is created out of morbidity and mortality without specific weighting. This is somewhat problematic as these are 2 very different outcomes

**Is it potentially problematic to have a 3 year time window for measurement if the time-period for comparison is shorter (i.e. if one is below average in year one, accrues some sort of penalty but then dramatically improves, one may continue to be regarded as below average for some time until the first year moves out of the 3 year window)? **

**The data elements are well defined except for a the time trend/year of surgery, which is a little unclear.

**Appropriate codes are included. **

**The calculation is clearly described. **

**Likely to be very consistently implemented. **

**It's not obvious why age, which is a continuous value, is/was broken up at 65 years with age functions, but maybe that is to better model the age relationship with outcome? **

**Simple yes/no for Cigarettes also seems like a coarse (but maybe convenient) grouping. **

2a2. Reliability Testing

Comments:

Reliability was tested and for those with even moderate volume, it is quite reliable. The only issue should be where lower volumes are concerned

**The reliability seems satisfactory for continued use as significant differences in performance are likely to be correctly identified. **

2b2. Validity Testing

Comments:

**Yes, but it is a manual process that it is uncertain whether this would be scalable. **

**I believe that the score is sufficiently valid such that reasonable conclusions about quality are possible based on one's score. **

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

this seems reasonable so long as the amount of missing data stays at a reasonably low level

**It wasn't clear if surgeons with few esophagectomies were excluded, but it seems like that would be sensible. **

**The risk adjustment strategy is reasonable. **

**The only variables that are not entirely clear is years of surgery/time trend. **

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.
- All data elements from participating institutions are submitted to the STS GTSD in an electronic format following a standard set of data specifications.
- STS GTSD participating institutions utilize data entry software products that are approved for the purposes of collecting and submitting STS GTSD data elements.
- There are no direct costs to collect data for this measure. STS General Thoracic Surgery Database participants (single surgeon or a group of surgeons) pay annual fees of $550 per surgeon for STS members and $700 per surgeon for non-STS members. In addition, there is a cost associated with purchasing data collection software which varies across vendors. STS GTSD participants have a separate agreement with their vendor, a process in which STS is not involved.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:
this is feasible, but can be time intensive. However current use of the software demonstrates it's feasibility
**The data elements are likely to be available, sometimes as discrete data elements, but probably they'll have to be abstracted for submission in many cases. **

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:

- In 2017, STS is planning to launch the general thoracic surgery component of STS Public Reporting Online. This is currently in the planning stage. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online
- The STS National Database is a Qualified Clinical Data Registry (QCDR). Prolonged length of stay is one of many measures that STS reports to CMS on behalf of consenting STS Adult Cardiac Surgery Database surgeons. For the STS General Thoracic Surgery Database (GTSD), physician quality reporting is in the planning stage.
- STS GTSD leaders and STS staff are reviewing general thoracic measures for inclusion in physician quality reporting.
- STS intends to include general thoracic measures in its 2017 QCDR self-nomination.

Improvement results:

- The developer provided the information below. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

	Total	Midwest	Northeast	South	West
# of participants	169 (100%)	35 (20.7%)	53 (31.4%)	54 (32.0%)	27 (16.0%)
# of patients	4557 (100%)	1325 (29.1%)	1499 (32.9%)	1176 (25.8%)	557 (12.2%)

Unexpected findings (positive or negative) during implementation:

- The developer reports no additional difficulties or unexpected findings or benefits, apart from those included throughout the submission form.

Potential harms:

- The developer reports no unintended consequence were noted.

Feedback :

- Measure has not been reviewed by MAP.

***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

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

it is not publicly reported. Data is fed back to the originating sites however and used.
**Looks like there is a good plan for public reporting, which presently isn't occurring. I'm not sure how prominent the proposed reporting measure is or if it would guide many patients. Seems possible that performance results might encourage low performance surgeons to aim for improvement or shift focus to other procedures. **

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0460
Measure Title: Risk-Adjusted Morbidity and Mortality for Esophagectomy for Cancer
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 3/14/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering 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:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
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

☒ Health outcome: discharge mortality or major morbidity
☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Measuring risk adjusted morbidity and mortality of patients undergoing esophagectomy for cancer provides surgeons and institutions the opportunity to evaluate outcomes and subsequently design quality improvement initiatives to address identified deficits. Utilization of the insight gained should promote improved patient outcome.

Esophageal cancer is an aggressive disease with a generally poor prognosis. The incidence of esophageal adenocarcinoma is increasing faster than any other malignancy in the United States. In 2015, there were an estimated 16,980 people diagnosed with esophageal cancer.

Esophagectomy, a relatively high morbidity and mortality operation, remains a key therapy in treating patients with localized esophageal cancer. The 30-day mortality rate following esophagectomy ranges between 2.7% and 11%. Within the STS GTSD, 24% of patients undergoing esophagectomy for cancer experienced major postoperative morbidity or death. Those with a major morbidity had a hospital discharge mortality of 11% while those patients without a major morbidity had a mortality rate of zero. This analysis identified a number of statistically significant predictors of major morbidity or mortality after esophagectomy for cancer. These factors included age, race, cardiac disease, impaired lung function, peripheral vascular disease, hypertension, diabetes, functional status, smoking status, and steroid use. Recognition of these predictors preoperatively, and modifying them when possible may provide improved outcomes, as measured by mortality, length of stay, postoperative quality of life, overall costs and resource utilization. Some of these preoperative predictors are modifiable, such as smoking status, and have been shown to reduce complication rates. Pulmonary complications are the major source of morbidity and mortality after esophageal resection, and numerous studies have identified various associated with these complications. Preoperative factors affecting pulmonary complications include advanced age, poor nutritional status, and poor cardiopulmonary reserve. Intraoperative factors associated with increased rates of pulmonary complications include increased blood loss, excessive fluid administration, prolonged operative times, advanced or proximal esophageal tumors, and more extensive operations, including the McKeown resection with three-field lymph node dissection. Postoperative factors associated with pulmonary complications include the development of atrial fibrillation, recurrent laryngeal nerve injury, and aspiration or other abnormality of deglutition. Enhanced recovery pathways and fast-track protocols, have been shown to reduce major morbidity and length of stay without increasing mortality or readmissions. Knowing their rate of risk adjusted morbidity and mortality after esophagectomy gives thoracic programs the opportunity to design quality improvement initiatives around deficiencies.

- Tomaszek S, Cassivi SD. Esophagectomy for the treatment of esophageal cancer. Gastroenterol Clin North Am. 2009 Mar;38(1):169-81.
- SEER database. March 8, 2016. Retrieved from http://seer.cancer.gov/statfacts/html/esoph.html
- Wright CD, Kucharczuk JC, O'Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.
- Kassis Edmund S, Kosinski Andrzej S, Ross Jr. Patrick, Koppes Katherine E, Donahue James M, Daniel Vincent C. Predictors of Anastomotic Leak After Esophagectomy: An Analysis of The Society of Thoracic Surgeons General

Thoracic Database. Ann Thorac Surg. 2013 Dec;96(6):1919-26. doi: 10.1016/j.athoracsur.2013.07.119. Epub 2013 Sep 24
- Schieman C, Wigle DA, Deschamps C, Nichols Iii FC, Cassivi SD, Shen KR, Allen MS. Patterns of operative mortality following esophagectomy. Dis Esophagus. 2012 Sep-Oct;25(7):645-51. doi: 10.1111/j.1442-2050.2011.01304.x. Epub 2012 Jan 13
- Hii MW, Smithers BM, Gotley DC, Thomas JM, Thomson I, Martin I, Barbour AP. Impact of postoperative morbidity on long-term survival after oesophagectomy. Br J Surg. 2013 Jan;100(1):95-104. doi: 10.1002/bjs.8973. Epub 2012 Nov 12
- Derogar M, Orsini N, Sadr-Azodi O, Lagergren P. Influence of major postoperative complications on health-related quality of life among long-term survivors of esophageal cancer surgery. J Clin Oncol. 2012 May 10;30(14):1615-9. doi: 10.1200/JCO.2011.40.3568. Epub 2012 Apr 2.
- Carrott PW, Markar SR, Kuppusamy MK, Traverso LW, Low DE. Accordion severity grading system: assessment of relationship between costs, length of hospital stay, and survival in patients with complications after esophagectomy for cancer. J Am Coll Surg. 2012 Sep;215(3):331-6. doi: 10.1016/j.jamcollsurg.2012.04.030. Epub 2012 Jun 8.
- Yoshida N, Baba Y, Hiyoshi Y, et al. Duration of Smoking Cessation and Postoperative Morbidity After Esophagectomy for Esophageal Cancer: How Long Should Patients Stop Smoking Before Surgery? World J Surg 2016 Jan;40(1):142-7.
- Atkins BZ, D'Amico TA. Respiratory complications after esophagectomy. Thorac Surg Clin 2006 Feb;16(1):35-48.
- Casado D, López F, Martí R. Perioperative fluid management and major respiratory complications in patients undergoing esophagectomy. Dis Esophagus. 2010 Sep;23(7):523-8.
- Markar SR, Karthikesalingam A, Low DE. Enhanced recovery pathways lead to an improvement in postoperative outcomes following esophagectomy: systematic review and pooled analysis. Dis Esophagus 2015 Jul;28(5):468-75.
- Shewale JB, Correa AM, Baker CM, et al. Impact of a Fast-track Esophagectomy Protocol on Esophageal Cancer Patient Outcomes and Hospital Charges. Ann Surg 2015 Jun;261(6):1114-23.

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

See response in 1a.2.

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

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

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

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

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

☐ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

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

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

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

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

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

☐ Yes → *complete section 1a.7*

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

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

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

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

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

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

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

Complete section 1a.7

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

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

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

Complete section 1a.7

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

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

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

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

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

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

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

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

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

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

_____

1a.8 OTHER SOURCE OF EVIDENCE

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

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

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

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1a._Evidence_-_0460_MM_for_Esophagectomy_for_Cancer.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
It is important for surgeons to be able to compare their surgical outcomes to those of peer institutions as a means of assessing results and improving quality of care. Measuring risk adjusted morbidity and mortality of patients undergoing esophagectomy for cancer provides surgeons and institutions the opportunity to evaluate outcomes and subsequently design quality improvement initiatives to address identified deficits. Utilization of the insight gained should promote improved patient outcome.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see the Appendix.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Please see the Appendix.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Esophageal cancer is an aggressive disease with a generally poor prognosis. The incidence of esophageal adenocarcinoma is increasing faster than any other malignancy in the United States. In 2015, there were an estimated 16,980 people diagnosed with esophageal cancer.

Esophagectomy, a relatively high morbidity and mortality operation, remains a key therapy in treating patients with localized esophageal cancer. The 30-day mortality rate following esophagectomy ranges between 2.7% and 11%. Within the STS GTSD, 24% of patients undergoing esophagectomy for cancer experienced major postoperative morbidity or death. Those with a major morbidity had a hospital discharge mortality of 11% while those patients without a major morbidity had a mortality rate of zero. This analysis identified a number of statistically significant predictors of major morbidity or mortality after esophagectomy for cancer. These factors included age, race, cardiac disease, impaired lung function, peripheral vascular disease, hypertension, diabetes, functional status, smoking status, and steroid use. Recognition of these predictors preoperatively, and modifying them when possible may provide improved outcomes, as measured by mortality, length of stay, postoperative quality of life, overall costs and resource utilization. Some of these preoperative predictors are modifiable, such as smoking status, and have been shown to reduce complication rates. Pulmonary complications are the major source of morbidity and mortality after esophageal resection, and numerous studies have identified various associated with these complications. Preoperative factors affecting pulmonary complications include advanced age, poor nutritional status, and poor cardiopulmonary reserve. Intraoperative factors associated with increased rates of pulmonary complications include increased blood loss, excessive fluid administration, prolonged operative times, advanced or proximal esophageal tumors, and more extensive operations, including the McKeown resection with three-field lymph node dissection. Postoperative factors associated with pulmonary complications include the development of atrial fibrillation, recurrent laryngeal nerve injury, and aspiration or other abnormality of deglutition. Enhanced recovery pathways and fast-track protocols, have been shown to reduce major morbidity and length of stay without increasing mortality or readmissions. Knowing their rate of risk-adjusted morbidity and mortality after esophagectomy gives thoracic programs the opportunity to design quality improvement initiatives around deficiencies.

1c.4. Citations for data demonstrating high priority provided in 1a.3

- Tomaszek S, Cassivi SD. Esophagectomy for the treatment of esophageal cancer. Gastroenterol Clin North Am. 2009 Mar;38(1):169-81.
- SEER database. March 8, 2016. Retrieved from http://seer.cancer.gov/statfacts/html/esoph.html .
- Wright CD, Kucharczuk JC, O´Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.
- Kassis Edmund S, Kosinski Andrzej S, Ross Jr. Patrick, Koppes Katherine E, Donahue James M, Daniel Vincent C. Predictors of Anastomotic Leak After Esophagectomy: An Analysis of The Society of Thoracic Surgeons General Thoracic Database. Ann Thorac Surg. 2013 Dec;96(6):1919-26. doi: 10.1016/j.athoracsur.2013.07.119. Epub 2013 Sep 24
- Schieman C, Wigle DA, Deschamps C, Nichols Iii FC, Cassivi SD, Shen KR, Allen MS. Patterns of operative mortality following esophagectomy. Dis Esophagus. 2012 Sep-Oct;25(7):645-51. doi: 10.1111/j.1442-2050.2011.01304.x. Epub 2012 Jan 13
- Hii MW, Smithers BM, Gotley DC, Thomas JM, Thomson I, Martin I, Barbour AP. Impact of postoperative morbidity on long-term survival after oesophagectomy. Br J Surg. 2013 Jan;100(1):95-104. doi: 10.1002/bjs.8973. Epub 2012 Nov 12
- Derogar M, Orsini N, Sadr-Azodi O, Lagergren P. Influence of major postoperative complications on health-related quality of life among long-term survivors of esophageal cancer surgery. J Clin Oncol. 2012 May 10;30(14):1615-9. doi: 10.1200/JCO.2011.40.3568. Epub 2012 Apr 2.
- Carrott PW, Markar SR, Kuppusamy MK, Traverso LW, Low DE. Accordion severity grading system: assessment of relationship between costs, length of hospital stay, and survival in patients with complications after esophagectomy for cancer. J Am Coll Surg. 2012 Sep;215(3):331-6. doi: 10.1016/j.jamcollsurg.2012.04.030. Epub 2012 Jun 8.
- Yoshida N, Baba Y, Hiyoshi Y, et al. Duration of Smoking Cessation and Postoperative Morbidity After Esophagectomy for Esophageal Cancer: How Long Should Patients Stop Smoking Before Surgery? World J Surg 2016 Jan;40(1):142-7.
- Atkins BZ, D´Amico TA. Respiratory complications after esophagectomy. Thorac Surg Clin 2006 Feb;16(1):35-48.
- Casado D, López F, Martí R. Perioperative fluid management and major respiratory complications in patients undergoing esophagectomy. Dis Esophagus. 2010 Sep;23(7):523-8.
- Markar SR, Karthikesalingam A, Low DE. Enhanced recovery pathways lead to an improvement in postoperative outcomes following esophagectomy: systematic review and pooled analysis. Dis Esophagus 2015 Jul;28(5):468-75.
- Shewale JB, Correa AM, Baker CM, et al. Impact of a Fast-track Esophagectomy Protocol on Esophageal Cancer Patient Outcomes and Hospital Charges. Ann Surg 2015 Jun;261(6):1114-23.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Lung, Esophageal, Surgery, Surgery : Thoracic Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Complications

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.)*
See Appendix

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:** S.15._Detailed_risk_model_specifications_-_0460_MM_for_Esophagectomy_for_Cancer.docx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The denominator time window was changed from 12 months to 36 months. The 36-month time window is necessary to obtain appropriate sample sizes for this measure.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer who developed any of the following postoperative conditions: bleeding requiring reoperation, anastomosis leak requiring medical or surgical treatment, reintubation, ventilation >48 hours, pneumonia, or discharge mortality.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator –
• Complications: During hospitalization regardless of length of stay or within 30 days of surgery if discharged
• Discharge mortality: during the same hospitalization as surgery regardless of timing
Denominator – 36 months

S.6. **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, 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.
Number of patients undergoing elective esophagectomy for esophageal cancer for whom:

1. Unexpected return to the operating room (ReturnOR - STS General Thoracic Surgery Database (GTSD), Version 2.2, sequence number 1720) is marked "yes" and primary reason for return to OR (ReturnORRsn – STS GTSD, Version 2.2, sequence number 1730) is marked "bleeding" or "anastomatic leak following esophageal surgery"

or

2. Postoperative events (POEvents - STS GTSD v 2.2, sequence number 1710) is marked "Yes" and one of the following items is marked:

a. Anastomotis requiring medical treatment only (i.e., interventional radiation drainage, NPO, antibiotics) (AnastoMed- STS GTSD v 2.2, sequence number 1950)
b. Reintubate, Reintube (STS GTSD v 2.2, sequence number 1850)
c. Initial ventilator support > 48 hours (Vent- STS GTSD v 2.2, sequence number 1840)
d. Pneumonia (Pneumonia- STS GTSD v 2.2, sequence number 1780)
e. Discharge Status (MtDCStat - STS GTSD v 2.2, sequence number 2200) is marked as "Dead"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of patients aged 18 years and older undergoing elective esophagectomy for esophageal cancer

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk : Individuals with multiple chronic conditions, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*

1. Esophageal cancer (EsophCancer- STS GTSD v 2.2, sequence number 1140) is marked "yes" and Category of Disease – Primary (CategoryPrim- STS GTSD v 2.2, sequence number 1300) is marked as one of the following:

(ICD-9, ICD-10)
Esophageal cancer, lower third(150.5, C15.5)
Esophageal cancer, middle third (150.4, C15.4)
Esophageal cancer, upper third (150.3, C15.3)
Malignant other part esophagus (150.8, C15.8)
Esophageal cancer, esophagogastric junction (cardia) (151.0, C16.0)

2. Primary procedure (Primary- STS GTSD v 2.2, sequence number 1500) is marked as one of the following:
Transhiatal-Total esophagectomy, without thoracotomy, with cervical esophagogastrostomy (43107)
Three hole-Total esophagectomy with thoracotomy; with cervical esophagogastrostomy (43112)
Ivor Lewis-Partial esophagectomy, distal two-thirds, with thoracotomy and separate abdominal incision (43117)
Thoracoabdominal-Partial esophagectomy, thoracoabdominal approach (43122)
Minimally invasive three hole esophagectomy (43XXX)
Minimally invasive esophagectomy, Ivor Lewis approach (43XXX)
Minimally invasive esophagectomy, Abdominal and neck approach (43XXX)
Total esophagectomy without thoracotomy; with colon interposition or small intestine reconstruction (43108)
Total esophagectomy with thoracotomy; with colon interposition or small intestine reconstruction (43113)
Partial esophagectomy, cervical, with free intestinal graft, including microvascular anastomosis (43116)
Partial esophagectomy, with thoracotomy and separate abdominal incision with colon interposition or small intestine (43118)
Partial esophagectomy, distal two-thirds, with thoracotomy only (43121)
Partial esophagectomy, thoracoabdominal with colon interposition or small intestine (43123)
Total or partial esophagectomy, without reconstruction with cervical esophagostomy (43124)

3. Status of operation (Status - STS General Thoracic Surgery Database v 2.2, sequence number 1420) is marked as "Elective"

4. Gender and discharge mortality status information are provided; Gender (STS GTSD v 2.2, sequence number 190) is marked as "Male" or "Female, and discharge status (MtDCStat- STS GTSD v 2.2, sequence number 2200) is marked as "Alive" or "Dead"

5. Only analyze first operation of hospitalization meeting criteria 1-4.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The model was developed by multivariate logistic regression. The details of risk adjustment model development were published in 2009:

Wright CD, Kucharczuk JC, O´Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.

The following covariates are included in the esophagectomy Morbidity/Mortality model:

Variable	Sequence #	Definition
Age	170	
Gender	190	
African-American Race	210	
Zubrod Score	820	Modeled as a linear trend
ASA Class	1470	Modeled as a linear trend
Congestive Heart Failure	540	
Coronary Artery Disease	550	
Peripheral Vascular Disease	560	
Insulin-dependent Diabetes	640, 650	If “Yes” for Diabetes (640) and “Insulin” marked for Diabetes Control (650)
Hypertension	520	
Steroid Use	530	
Renal Dysfunction	680	
Cigarette Smoking (Ever)	730	
Body Mass Index	490, 500	Calculated using height and weight values – kg/m2. Modeled as a linear trend.
Preoperative Therapy	580, 600	If “Yes” for Preoperative chemotherapy (580) or “Yes” for Preoperative Thoracic Radiation Therapy (600)

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion

If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
To maximize use of available data, when encountering records with missing values of model covariates (with the exception of age and gender), the missing values were imputed. Patient records missing age or gender were excluded. Missing values of binary (yes/no) risk factors were conservatively imputed to the negative condition. Remaining variables were imputed to the median (BMI) or mode (race, Zubrod score, ASA class).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS General Thoracic Surgery Database (GTSD) Version 2.2; STS GTSD Version 2.3 went live on January 1, 2015.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form

2.1_Testing_-_0460_MM_for_Esophagectomy_for_Cancer.4.5.16.docx,2.1_Testing_-_0460_MM_for_Esophagectomy_for_Cancer.4.12.16.docx

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0460	**NQF Project:** Clinician Level Perioperative Care

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
STS General Thoracic Surgery Database - Compare results between two consecutive 6-month time intervals during 2009: January 2009 - June 2009 and July 2009 - December 2009

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
To assess temporal reliability of the proposed measure, we estimated hospital rates of mortality or major morbidity for two consecutive 6-month time intervals during 2009 and compared the results. Only 12 months of data were available for the current data version v2.081. Thus, we were unable to consider time intervals longer than 6 months. Only hospitals with data for both time periods were included. Hospital estimates for each time period were estimated using hierarchical models, as described above. The correlation between hospital estimates in two time periods was assessed graphically and summarized by the Pearson correlation coefficient.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The Pearson correlation between hospital-specific rates of mortality or major morbidity in the 1st vs. 2nd half of 2009 was 0.50.

Site-specific proportions of patients experiencing morbidity/mortality endpoint (based on hierarchical model)

Correlation between performance in two consecutive time periods

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
STS General Thoracic Surgery Database

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
When data arrive at the data warehouse, they are checked carefully for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report

that is generated automatically following each harvest file submission. Upon receipt of the data quality report, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis. If the data inconsistencies are not changed by the participant prior to harvest close, the data warehouse performs consistency edits and/or parent/child edits on the data in order for them to be analyzable. Participants are informed of such edits to their data in the Data Quality Report.

Since 2006, the Iowa Foundation for Medical Care (IFMC) has conducted audits of the STS Adult Cardiac Surgery Database on the Society´s behalf. Beginning in the fall of 2010, IFMC will conduct audits of the STS General Thoracic Surgery Database to evaluate the accuracy, consistency and comprehensiveness of data collection which will validate the integrity of the data. 5% of participants will be randomly selected annually. Auditors will validate case inclusion and twenty lobectomy cases will be randomly chosen for review of thirty individual data elements. The auditors will abstract each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates will be calculated for each of the 30 elements as well as an overall agreement rate.

In addition, validity was confirmed and is regularly assessed by an expert panel of thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Task Force on Quality Initiatives and the STS Workforce on National Databases.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Please see Risk Adjustment Type section above

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Detailed information regarding the risk adjustment model can be found in the attachment:

Wright CD, Kucharczuk JC, O´Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Data for this analysis were based on 791 patients who underwent surgery during 2009 at hospitals participating in the STS General Thoracic Surgery Database and met the inclusion criteria for the measure. Only hospitals submitting operations in both semesters of 2009 were included.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
For each patient, we determined whether or not at least one of the endpoints of the composite endpoint (discharge mortality, bleeding requiring reoperation, anastomosis requiring medical treatment, anastomosis requiring surgical treatment, reintubation, ventilation >48 hours) occurred. Hospital-specific probabilities of the composite endpoint were estimated in a Bayesian hierarchical model. These model-based estimates were used to control variation due to random statistical fluctuations while estimating true signal variation. Hospital-specific estimates were summarized by percentiles and plotted along with 95% credible intervals to illustrate between-hospital variation.

Results for each participant are presented in two forms: (1) the estimated risk-adjusted rate (RAR); and (2) the estimated standardized incidence ratio (SIR). The RAR is interpreted as the outcome rate that would be observed hypothetically for a participant if the participant performed surgery on each eligible patient in the STS General Thoracic Surgery Database. This hypothetical quantity cannot be observed directly but may be estimated in a statistical model, as described below.

A participant's SIR is defined as the ratio of the participant's RAR divided by the overall STS observed outcome rate.
SIR of participant = RAR of participant / overall STS observed rate.

An SIR value greater than 1.0 implies that the participant's risk-adjusted outcome rate is higher than the overall STS observed rate. Conversely, an SIR value less than 1.0 implies that the participant's risk-adjusted outcome rate is lower than the overall STS observed rate.

To account for uncertainty in the estimation of RAR and SIR, the estimates of these quantities are accompanied by 95% credible intervals (CI). The 95% CI indicates the range of RAR and SIR values that are plausible in light of the observed data. If the 95% CI for a participant's SIR includes the null value 1.0, then we cannot reliably distinguish this participant's performance from the STS average - either the participant's performance was close to average or else the participant's sample size was too small to make a reliable determination.

Statistical Model
Random effects logistic regression models were used to compare each participant's event rate in a manner that adjusts for case mix and accounts for uncertainty due to small sample sizes. Random effects models use data from all database participants when estimating the event rate of a single participant, thereby borrowing strength to obtain a more reliable estimate. Each outcome was adjusted for its own set of patient
factors (listed below) and included a separate random effect parameter for each participant in the analysis. The model has the form:

where pji denotes the probability of prolonged stay for the i-th patient at participant j; ej denotes a (random effect) intercept parameter for participant j; and xjiq denotes the value of q-th covariate for the i-th patient at the j-th participant. The terms xjiq represent quantitative risk factors such as age and FEV; and binary indicator variables (e.g. 1=male, 0=female).

Parameters of the random effects logistic model were estimated in a Bayesian framework using WinBUGS software. Unlike conventional statistical methods, the results of Bayesian analyses are expressed in terms of probabilities. For example, we might be 95% sure that a participant has a better-than-average risk-adjusted rate. The Bayesian 95% credible interval (CI) has the following interpretation. In light of the observed data, it is 95% likely that the true value of the parameter of interest lies in the indicated interval.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

The endpoint of discharge mortality or major morbidity occurred in 26.7% of eligible patients. Hospital-specific estimates ranged from 18.9% to 44.1%.

Distribution of hospital-specific estimated probabilities of mortality or morbidity

Min	25th percentile	Median	75th percentile	Maximum
18.9%	25.4%	27.1%	30.2%	44.1%

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:**

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**

2.1-2.3 **Supplemental Testing Methodology Information:**

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

UPDATED MEASURE TESTING

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0460

Measure Title: Risk-Adjusted Morbidity and Mortality for Esophagectomy for Cancer

Date of Submission: 3/14/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in**

performance;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.
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.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*).

STS General Thoracic Surgery Database (GTSD) Version 2.2

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

July 1, 2012 – June 30, 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.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: 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*)

The calculation of this measure using the 36 months from July 1, 2012 to June 30, 2015 included 4,557 operations from 169 STS General Thoracic Surgery Database participants. Description of the distribution of participant size (patient volume) overall and by geographic region is below

	Total	**Midwest**	**Northeast**	**South**	**West**
# of participants	169	35	53	54	27
# of patients	4557	1325	1499	1176	557

Participant size (volume)

Mean	27	37.9	28.3	21.8	20.6
SD	39.1	53.7	38.8	33.3	23.4
IQR	31	41.5	35	21.8	26
Minimum	1	1	1	1	1
10% percentile	2	1	1	2	2
20% percentile	3	3	3	4	3.2
30% percentile	5	5	5	5	4.8
40% percentile	8	12	8	9	7.4
Median	13	23	12	12	13
60% percentile	19	30	21	14	14.6
70% percentile	29	43	33	18	24
80% percentile	42	47	43	31	33.6
90% percentile	61	87	69	52	50.4
Maximum	259	259	173	214	83

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*)

Includes 4,557 eligible patients. Patient characteristics are below.

Age (years), mean (SD)	63.5 (9.5)
Female	17.6%
Black race	3.4%
Congestive heart failure	2.4%
Coronary artery disease	18.7%
Peripheral vascular disease	4.1%
Zubrod score	
0	23.1%
1	72.4%
2-5	4.5%
ASA Class	
I-II	14.2%
III	76.7%
IV-V	9.0%
Insulin diabetes	5.7%
Hypertension	56.0%
Steroids	1.6%
Renal dysfunction	1.0%
Induction therapy	69.1%
Cigarette use	72.7%
Body Mass Index (kg/m2), mean, (SD)	27.8 (6.0)
Year of surgery	
2012	16.2%
2013	34.1%
2014	34.3%
2015	15.4%

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.

No data differences are present.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

No SDS variables were collected and utilized.

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 is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. Accordingly, reliability was defined as the square of the Pearson correlation coefficient (ρ^2) between the set of participant-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$ and the corresponding unknown true values $\theta_1, \ldots, \theta_N$ (N=number of participants), that is:

$$\rho^2 = \frac{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)\left(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h\right)}{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)^2 \sum_{j=1}^{N}\left(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h\right)^2}.$$

The quantity ρ^2 was estimated by its posterior mean estimated from a MCMC chain of length 2000, namely, $\hat{\rho}^2 = \frac{1}{2000}\sum_{l=1}^{2000}\rho_{(l)}^2$ where

$$\rho_{(l)}^2 = \frac{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)\left(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)}\right)}{\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h\right)^2 \sum_{j=1}^{N}\left(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)}\right)^2}$$

with $\theta_j^{(l)}$ denoting the value of θ_j on the l-th MCMC sample and $\hat{\theta}_j = \frac{1}{2000}\sum_{l=1}^{2000}\theta_j^{(l)}$ denoting the estimated posterior mean of θ_j. A 95% credible interval for ρ^2 was obtained by calculating the 50-th smallest and 50-th largest values of $\rho_{(l)}^2$ across the 2000 MCMC samples ($l = 1, \ldots, 2000$).

Kozower BD, O'Brien SM, Kosinski AS, Magee MJ, Dokholyan R, Jacobs JP, Shahian DM, Wright CD, Fernandez FG. The Society of Thoracic Surgeons Composite Score for Rating Program Performance for Lobectomy for Lung Cancer. Ann Thorac Surg. 2016 Jan 16. pii: S0003-4975(15)01753-1. doi: 10.1016/j.athoracsur.2015.10.081. [Epub ahead of print]. PMID: 26785936.

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*)

Based on all the 169 participants the reliability (proportion of signal variation) is 44.4%, 95% credible interval [CrI] (34.0%, 53.9%). Reliability increases when considering participants with a particular minimum number of procedures per year as displayed below.

	All participants	≥5 procedures per year	≥10 procedures per year	≥15 procedures per year	≥20 procedures per year
Number of participants	169	75	50	29	20
Reliability	44.4%	67.9%	71.1%	72.5%	80.6%
95% CrI for reliability	(34%, 53.8%)	(55.7%, 77.9%)	(57.8%, 81.8%)	(56.9%, 84.9%)	(65.5%, 91.1%)

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 improves with volume and it is comparable to the reliability for other measures for STS GTSD.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

Critical data elements

When data arrive at the data warehouse, they are checked for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report (DQR) that is generated automatically following each harvest file submission. Upon receipt of the DQR, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis.

In addition, participating sites are randomly selected for participation in the STS General Thoracic Surgery Database (GTSD) Audit, which is designed to evaluate the accuracy, consistency, and comprehensiveness of data collection and ultimately validate the integrity of the data contained in the database. Telligen, formerly the Iowa Foundation for Medical Care, has conducted audits on behalf of STS since 2006. In 2015, ten percent of randomly selected STS GTSD participants (N = 25, an increase from 24 in 2014 and 18 in 2013) were audited. The audit process involves re-abstraction of data for 20 cases records (at least 15 lobectomy and up to 5 esophagectomy) and comparison of 40 STS GTSD V2.2 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each variable, each variable category and overall. In 2015, the overall aggregate agreement rate was 97.02%, demonstrating that the data contained in the STS GTSD are both comprehensive and highly accurate.

Data Analysis

Aggregate agreement rates were computed for all facilities by calculation of the sum of all facilities' numerators divided by the sum of all facilities' denominators, for each individual variable, each variable category and overall.

Chi-square statistics were calculated to identify any possible relationships between the data collection process variables and agreement rates. Tests where the chi-square statistic had a probability of less than 5% ($p<0.05$) were considered to show statistically significant differences in agreement rate between the levels of the process measure.

Agreement Rate Results

Database validity was evaluated by re-abstraction of defined variables from the medical records and comparison to submitted data. Agreement rates were calculated at the individual variable level, category level and overall. Aggregate agreement rates are presented in the table below.

CATEGORY	*FIELD_NAME*	*NUM*	*DEN*	*Agreement Rate*
PRE-OPERATIVE EVALUATION	OVERALL_ALL FIELDS	6213	6452	96.30%
Pre-Operative Evaluation	Admission Date	496	500	99.20%
Pre-Operative Evaluation	Prior Cardiothoracic Surgery	489	500	97.80%
Pre-Operative Evaluation	Pre-Op Chemo-Current Malignancy	494	500	98.80%
Pre-Operative Evaluation	Pre-Op Thoracic Radiation Therapy	495	500	99.00%
Pre-Operative Evaluation	Diabetes	409	415	98.55%
Pre-Operative Evaluation	Diabetes Control	69	82	84.15%
Pre-Operative Evaluation	Cigarette Smoking	475	500	95.00%
Pre-Operative Evaluation	Pulmonary Function Tests Performed	394	415	94.94%

CATEGORY	*FIELD_NAME*	*NUM*	*DEN*	*Agreement Rate*
Pre-Operative Evaluation	FEV1 Predicted	349	377	92.57%
Pre-Operative Evaluation	Zubrod Score	467	499	93.59%
Pre-Operative Evaluation	Lung Cancer	408	409	99.76%
Pre-Operative Evaluation	Clinical Staging Method-Lung-EBUS	355	366	96.99%
Pre-Operative Evaluation	Clinical Staging Method-Lung-PET or PET/CT	337	366	92.08%
Pre-Operative Evaluation	Lung Cancer Tumor Size	338	364	92.86%
Pre-Operative Evaluation	Lung Cancer Nodes	354	361	98.06%
Pre-Operative Evaluation	Esophageal Cancer	85	85	100.0%
Pre-Operative Evaluation	Clinical Staging Method-Esophageal-EUS	62	71	87.32%
Pre-Operative Evaluation	Esophageal Cancer Tumor	70	71	98.59%
Pre-Operative Evaluation	Esophageal Cancer Nodes	67	71	94.37%
Diagnosis And Procedures	OVERALL_ALL FIELDS	4663	4809	96.96%
Diagnosis And Procedures	Category of Disease-Primary	493	500	98.60%
Diagnosis And Procedures	Date of Surgery	500	500	100.0%
Diagnosis And Procedures	Procedure Start Time	481	500	96.20%
Diagnosis And Procedures	Procedure End Time	463	500	92.60%
Diagnosis And Procedures	ASA Classification	481	500	96.20%
Diagnosis And Procedures	Procedure	493	500	98.60%
Diagnosis And Procedures	Patient Disposition	481	500	96.20%
Diagnosis And Procedures	Pathological Staging-Lung Cancer-T	355	359	98.89%
Diagnosis And Procedures	Pathological Staging-Lung Cancer-N	357	359	99.44%
Diagnosis And Procedures	Lung Cancer - Number of Nodes	339	359	94.43%
Diagnosis And Procedures	Pathological Staging-Esophageal Cancer-T	86	90	95.56%
Diagnosis And Procedures	Pathological Staging-Esophageal Cancer-N	69	71	97.18%
Diagnosis And Procedures	Esophageal Cancer-Number of Nodes	65	71	91.55%
POST-OPERATIVE EVENTS	OVERALL_ALL FIELDS	1586	1598	99.25%
Post-Operative Events	Return to OR	496	500	99.20%
Post-Operative Events	Pneumonia	494	500	98.80%
Post-Operative Events	Initial Vent Support >48 Hours	498	500	99.60%
Post-Operative Events	Gastric Outlet	98	98	100.0%
DISCHARGE	OVERALL_ALL FIELDS	1950	1995	97.74%
Discharge	Discharge Date	500	500	100.0%
Discharge	Discharge Status	494	500	98.80%
Discharge	Readmission within 30 Days of Discharge	486	495	98.18%
Discharge	Status at 30 Days	470	500	94.00%
	OVERALL_ALL FIELDS	14412	14854	97.02%

There were 14,854 total variables abstracted and of those 14,412 variables matched, resulting in an overall agreement rate of 97.02%.

Process Variable Correlation Tables

The relationships between process variables and overall agreement rates were examined and included:

- Facility data collection performed from electronic medical records or a combination of paper and electronic medical records and overall agreement rate
- Facility data collection method (concurrent/retrospective/both) and overall agreement rate
- Data collection performed by a single abstractor or multiple staff and overall agreement rate
- Attendance at the annual data managers' meeting, STS Advances in Quality and Outcomes (AQO) Conference, and overall agreement rate
- Agreed upon abstraction location for data elements documented in multiple locations and overall agreement rate

Relationship between Data Collection Source & Agreement Rate
Facilities using an electronic health record (EHR) for data collection had higher agreement rates, 97.36%, than those facilities using both paper medical records and an EHR, 96.31%. There were no facilities that used paper medical records alone (p<0.0004).

Relationship between Data Collection Method & Agreement Rate
Facilities collecting data retrospectively have higher agreement rates, 97.55%, than those facilities collecting data concurrently, 96.18%, or both, 96.38% (p=<0.0001).

Relationship between Data Collection Performed By & Agreement Rate
Facilities with a single individual performing data abstraction have higher agreement rates, 98.02%, than those facilities that have multiple individuals performing data abstraction, 96.24% (p<0.0001).

Relationship between Attendance at AQO Conference & Agreement Rate
Facilities having staff attend the annual AQO Conference have higher agreement rates, 97.25%, than those that do not have staff attend, 96.11% (p<0.0012).

Relationship between Have an Agreed Upon Location & Agreement Rate
Facilities that utilize an agreed upon location for data elements recorded in multiple locations have higher agreement rates, 97.31%, than facilities that do not utilize an agreed upon location, 93.61% (p<0.0001).

In addition, validity is regularly assessed by an expert panel of general thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Quality Measurement Task Force, and the STS Task Force on Quality Initiatives, all of which report to the STS Workforce on National Databases.

Performance measure score
The measure is regarded as useful and valid by its intended users. The measure was developed with a panel of surgeon experts and statisticians. Differences in the measure across participants are clinically meaningful.

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

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

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

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

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

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

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with** 16 **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Covariates for the model were selected a-priori based on clinical relevance, literature review, and empirical analysis by a panel of physicians and statisticians. The details of model development were published in 2009 by Wright and colleagues. Risk of mortality and other short-term clinical outcomes is mostly influenced by clinical factors present on admission, such as age, comorbidities, cancer stage, and prior therapy. By convention, given the plausible causal pathways leading to these outcomes, risk models used for mortality profiling have generally excluded non-clinical patient factors or local environmental factors, as their inclusion might theoretically adjust out important inequities in care.

Disparities in postoperative morbidity and mortality based on socioeconomic status (SES) have been demonstrated after major cancer surgery. There is mixed evidence in the literature (see below) linking SES to poorer outcomes after esophagectomy, in particular. Given the mixed evidence supporting SES as a risk factor, it was not included in the model.

Reames et al demonstrated that patients in the lowest quintile of SES have significantly increased rates of failure to rescue from a major complication after undergoing esophagectomy, pancreatectomy, partial or total gastrectomy, colectomy, lung resection, or cystectomy for cancer. However, when controlling for patient and hospital factors, SES for esophagectomy was no longer a predictor of failure to rescue in this cohort.

Sammon et al took Patients undergoing 1 of 8 major cancer surgeries (colectomy, cystectomy, esophagectomy, gastrectomy, hysterectomy, lung resection, pancreatectomy, and prostatectomy) within the Nationwide Inpatient Sample from 1999-2009. Logistic regression models fitted with generalized estimating equations were used to estimate primary predictors (procedure, age, gender, race, insurance, Charlson Comorbidity Index, hospital volume, and hospital bed size) effect on sepsis and sepsis-associated mortality. They found that the odds of sepsis were highest among esophagectomy patients (odds ratio [OR]: 3.13, 2.76-3.55) and those with non-private insurance (OR: 1.33, 1.19-1.48 to OR: 1.89, 1.71-2.09).

Morgan et al examined a total of 1196 consecutive patients with esophageal carcinoma presenting to a regional multidisciplinary team between 1 January 1998 and 31 August 2005 and deprivation scores were calculated using the Indices of Multiple Deprivation (IMD) of the National Assembly for Wales, UK. Stage of disease and morbidity did not correlate with deprivation quintile, but operative mortality was non-statistically significantly greater in quintile 1 versus 5 (1.9% versus 5.8%, p = 0.281). Overall 5-year survival for those patients undergoing esophagectomy was unrelated to deprivation quintile (1 versus 5, 24% versus 33%, p = 0.8246), but was lower following definitive chemoradiotherapy (dCRT) for the least deprived quintiles (1, 2 & 3 versus 4 & 5, 35% versus 16%, p = 0.0272).

- Wright CD, Kucharczuk JC, O'Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.
- Reames B, Birkmeyer N, Dimick J, et al. Socioeconomic Disparities in Mortality After Cancer Surgery: Failure to Rescue. JAMA Surg. 2014;149(5):475-481.
- Sammon JD, Klett DE, Sood A. Sepsis after major cancer surgery. J Surg Res. 2015 Feb;193(2):788-94.
- Morgan MA, Lewis WG, Chan DS, et al. Influence of socio-economic deprivation on outcomes for patients diagnosed with oesophageal cancer. Scand J Gastroenterol. 2007 Oct;42(10):1230-7.

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

Mortality or Major Mortality Model		
Variable	**OR (95% CI)**	**p**
Age (per 1 yr increase below 65 yrs)	0.993 (0.980, 1.006)	0.28
Age (per 1 yr increase above 65 yrs)	1.023 (1.007, 1.039)	0.006
Female	0.916 (0.765, 1.097)	0.34
Black race	1.373 (0.963, 1.957)	0.085
Congestive heart failure	2.367 (1.586, 3.531)	<0.001
Coronary artery disease	1.147 (0.961, 1.369)	0.13
Peripheral vascular disease	1.267 (0.923, 1.740)	0.15
Zubrod score (vs. 0)		<0.001
1	1.010 (0.860, 1.188)	
>1	1.886 (1.373, 2.592)	
ASA risk class (vs. I or II)		0.015
III	1.065 (0.869, 1.305)	
IV or V	1.464 (1.099, 1.949)	
Insulin diabetes	1.259 (0.959, 1.652)	0.10
Hypertension	1.150 (0.993, 1.332)	0.062
Steroids	1.656 (1.018, 2.694)	0.046
Renal dysfunction	1.342 (0.729, 2.470)	0.35
Induction therapy	1.002 (0.866, 1.160)	0.97
Cigarette use	1.401 (1.196, 1.642)	<0.001
Body mass index (kg/m2) (per 1 unit)	1.018 (1.007, 1.029)	0.002
Year of surgery (per 1 yr increase)	0.970 (0.903, 1.042)	0.40

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)*

No sociodemographic (SDS) variables were collected and utilized.

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

The model was assessed for discrimination by means of C-index and for goodness-of-fit through Hosmer-Lemeshow statistic.

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

If stratified, skip to 2b4.9

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

C-statistic is 0.614.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.62 (Chi-Square=6.27, df=8).

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

Not provided

2b4.9. Results of Risk Stratification Analysis:

N/A

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

The results demonstrated that the risk model is well calibrated and has good discrimination power. It is suitable for controlling differences in case-mix between participants.

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

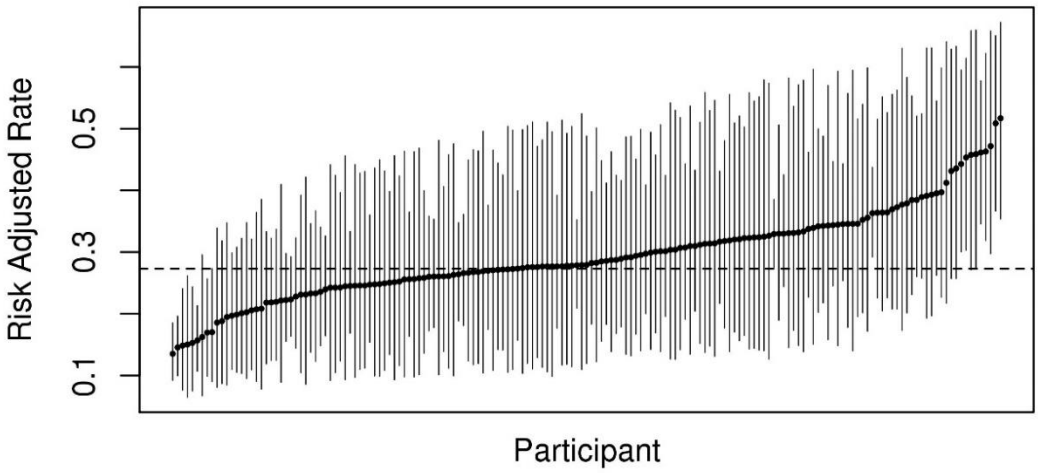
2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

Participant-specific risk-adjusted rates (RAR) of the endpoint were estimated within a Bayesian hierarchical logistic regression model. Participant-specific RARs were plotted along with corresponding 95% credible intervals to illustrate the between-participant variation.

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

Better than expected and worse than expected participants are distinguishable because they have the 95% credible intervals below or above the STS average as seen on the plot below.



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

The identified differences in performance are clinically meaningful and allow differentiation between participants based on estimates and 95% credible intervals.

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

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

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

2b6.3. 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*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

The quality of data in STS General Thoracic Surgery Database has been improving. We managed the missing data with imputation. Missing body mass index (BMI) values were imputed utilizing median of the observed BMI values. For binary risk factors, missing values were considered as indicating absence of the risk factor.

2b7.2. 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*)

Variables with missing data were: congestive heart failure (4.59%), coronary artery disease (4.43%), peripheral vascular disease (4.63%), insulin diabetes (5.24%), hypertension (4.32%), steroid use (7.83%), BMI = 4.63%, induction therapy (4.08%), renal dysfunction (7.22%). Other covariates had no missing data.

2b7.3. 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*)

The missing data approach was compared, as a sensitivity analysis, to multiple imputations and the results were insensitive to the approach. Thus, the simpler approach was taken as explained in the reference below.

Wright CD, Kucharczuk JC, O'Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg. 2009 Mar;137(3):587-95.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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.**

The STS General Thoracic Surgery Database (GTSD) has more than 270 participants, and local availability of data elements in electronic format varies across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS GTSD in an electronic format following a standard set of data specifications. STS GTSD participating institutions utilize data entry software products that are approved for the purposes of collecting and submitting STS GTSD data elements.

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The data elements included in this measure have been standard in the STS General Thoracic Surgery Database for at least 3 years and some of them have been part of the database for more than 15 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care. Every 3 years, the STS GTSD undergoes a specification upgrade (i.e., a process including surgeon leadership, staff, database managers, and programmers to review and update data fields and their respective definitions to ensure they reflect current practice).

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)*.

Data Collection:

There are no direct costs to collect data for this measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon leaders' time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS General Thoracic Surgery Database participants (single surgeon or a group of surgeons) pay annual fees of $550 per surgeon for STS members and $700 per surgeon for non-STS members. In addition, there is a cost associated with purchasing data collection software which varies across vendors. STS GTSD participants have a separate agreement with their vendor, a process in which STS is not involved.

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.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) STS General Thoracic Surgery Database – 273 Participants http://www.sts.org/national-database/database-managers/general-thoracic-surgery-databas Quality Improvement (Internal to the specific organization) STS General Thoracic Surgery Database – 273 Participants http://www.sts.org/national-database/database-managers/general-thoracic-surgery-databas

4a.1. For each CURRENT use, checked above, provide:

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

Please see above.

4a.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?*)

In 2017, STS is planning to launch the general thoracic surgery component of STS Public Reporting Online. This is currently in the planning stage. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online

The STS National Database is a Qualified Clinical Data Registry (QCDR). Prolonged length of stay is one of many measures that STS reports to CMS on behalf of consenting STS Adult Cardiac Surgery Database surgeons. For the STS General Thoracic Surgery Database (GTSD), physician quality reporting is in the planning stage. STS GTSD leaders and STS staff are reviewing general thoracic measures for inclusion in physician quality reporting. STS intends to include general thoracic measures in its 2017 QCDR self-nomination.

4a.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.*)

Please see 4a.2.

4b. 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.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Please see 1b.2 and 1b.4

The performance of each participant is calculated twice a year using an updated 3-year time window. The performance is compared to the average STS performance within the considered 3-year time window. These results are shared with all participants through a semiannual feedback report.

	Total	Midwest	Northeast	South	West
# of participants	169 (100%)	35 (20.7%)	53 (31.4%)	54 (32.0%)	27 (16.0%)
# of patients	4557 (100%)	1325 (29.1%)	1499 (32.9%)	1176 (25.8%)	557 (12.2%)

4b.2. 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.
N/A

4c. 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).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
We are not aware of any negative unintended consequences.

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

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 completely harmonized?

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

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:** Appendix_-__0460_MM_for_Esophagectomy_for_Cancer.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

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.

- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives (TQI); surgeon leader/clinical expert in adult cardiac surgery
- Benjamin Kozower, MD – Chair, General Thoracic Surgery Database Task Force; surgeon leader/clinical expert in general thoracic surgery
- William Burfeind, MD – Surgeon leader/clinical expert in general thoracic surgery
- Robert Welsh, MD – Surgeon leader/clinical expert in general thoracic surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Felix Fernandez, MD – Surgeon leader/clinical expert in general thoracic surgery
- Cameron Wright, MD – Surgeon leader/clinical expert in general thoracic surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Andrzej Kosinski, PhD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, Senior Manager of the STS National Database and Patient Safety

Members of the STS TQI and the GTSDTF provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

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:** 03, 2016

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?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** None

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 #: 0508
Measure Title: Diagnostic Imaging: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms
Measure Steward: American College of Radiology (ACR)
Brief Description of Measure: Percentage of final reports for screening mammograms that are classified as "probably benign"
Developer Rationale: The "probably benign" assessment category is reserved for findings that have a high probability (=98%) chance of being benign and should not be used as a category for indeterminate findings. Inappropriate designation of findings as "probably benign" can result in unnecessary follow-up of lesions that could have been quickly classified or delayed diagnosis and treatment of cancerous lesions. Published guidance documents emphasize the need to conduct a complete diagnostic imaging evaluation before making a probably benign (Category 3 assessment; making it inadvisable to use the probably benign categorization when interpreting a screening mammogram. Immediate completion of a diagnostic imaging evaluation for abnormal screening mammograms eliminates potential anxiety that women would endure with the short interval follow-up that is recommended for "probably benign" findings. The "probably benign" assessment category is reserved for findings that have a high probability (=98%) chance of being benign and should not be used as a category for indeterminate findings. Inappropriate designation of findings as "probably benign" can result in unnecessary follow-up of lesions that could have been quickly classified or delayed diagnosis and treatment of cancerous lesions. Published guidance documents emphasize the need to conduct a complete diagnostic imaging evaluation before making a probably benign (Category 3 assessment; making it inadvisable to use the probably benign categorization when interpreting a screening mammogram. Immediate completion of a diagnostic imaging evaluation for abnormal screening mammograms eliminates potential anxiety that women would endure with the short interval follow-up that is recommended for "probably benign" findings.

Numerator Statement: Final reports classified as "probably benign"
Denominator Statement: All final reports for screening mammograms
Denominator Exclusions: No Denominator Exclusions or Denominator Exceptions

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Registry
Level of Analysis: Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Oct 24, 2008

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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 2008

- The evidence for this measure was based on the guideline recommendation from the American College of Radiology (ACR) Breast Imaging Reporting and Data System (BI-RADS®) Atlas, 2003: Do not use 'probably benign' (Category 3) in interpreting screening examinations. The strength of evidence was not ranked.

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 following updates:

- The developer provided a brief statement describing the measure focus and the health outcome: The practice of rendering category 3 assessments directly from screening examination also has been shown to result in adverse health outcomes: 1) unnecessary follow-up of many lesions that could have been promptly assessed as benign, and 2) delayed diagnosis of a small number of cancers that otherwise may have been smaller in size and less likely to be advanced in stage.
- The developer provided one guideline from the American College of Radiology (ACR) for the performance of screening and diagnostic mammography:
    - Overall final assessment of findings should be based on all imaging studies performed up to that day. In addition, they must be classified according to the FDA-approved final assessment categories and should follow the categories defined in the ACR BI-RADS® 5th edition, 2012 (or any subsequent revisions). The BI-RADS® provides a framework for reporting, lesion assessment, imaging-pathologic correlation, quality improvement, and medical outcomes auditing. **Level of evidence: Not graded**
- The developer also provided a recommendation from the USPSTF: Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation:
    - This recommendation includes biennial screening mammography for women within different age groups and risks. **Level of evidence:** Varies depending on the recommendation.
- The developer states that another source of evidence for this measure was the ACR Breast Imaging Reporting and Data System (BI-RADS®) Atlas, a quality assurance tool used to standardize reporting, bring clarity to breast imaging interpretations and management recommendations, and to facilitate outcome monitoring.
- The developer also provides eight studies from the literature addressing the 'probably benign' category.
- The developer did not provide empirical evidence to support this process measure. The ACR guideline does not address the 'probably benign' category for screening mammograms and the level of evidence is not graded. The USPSTF recommendation focuses on mammography screening overall but does not provide evidence-based recommendations for interpreting results.

Exception to evidence – In the absence of empirical evidence to support this process measure, the Committee may consider an exception to the evidence requirement with adequate justification.

Guidance from the Evidence Algorithm

Process measure/no systematic review/guidelines submitted do not align with the measure focus (Box 3)→no empiric evidence (Box 7)→ INSUFFICIENT – Committee to determine whether an exception is justified.

Questions for the Committee:

- *Are you aware of any evidence linking the use of the 'probably benign' category to health outcomes?*
- *For possible exception to the evidence criterion:*
  - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
  - *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
  - *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable for this process of care without empirical evidence?*

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

Rationale: Lack of empirical evidence to support the process of care

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 the following data for current performance for 45,558 physicians participating in the CMS Physician Quality Reporting System:

	2012-2014	2012	2013	2014
Performance rate	2.74%	2.09%	5.48%	0.49%
Range	25th percentile: 0.33% 50th percentile: 0.74% 75th percentile: 1.96%	--	--	--
Reporting rate of eligible professionals (PQRS)	66.77%	51.78%	64.95%	91.73%

- The developer also states that based on claims data the use of code 3 (probably benign) on screening mammograms is 0.49%; the median recall rate in Hospital Compare is approximately 8%. The registry rate of code 3 (probably benign) on screening mammography is 1.56% and the minimally acceptable maximum recall rate standard is 12%.
- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).

Disparities:

- The developer did not provide disparities data from the measure as specified – this is required for endorsement maintenance.
- The developer did not provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement as required when data on disparities from the measure as specified is not provided.

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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 to Support Measure Focus

Comments:

Yes. The appropriate use of BI-RADS 3 only after full diagnostic evaluation is supported by evidence.

1b. Performance Gap

Comments:

Yes. This is a renewal. Although the performance is very high (98.4%), that actually leaves room for improvement. For every 1000 mammograms read, that means 16 patients were given BI-RADS 3 from screening, which is still too high.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data sources: Clinical Electronic claims and registry data

Specifications:

- The level of analysis is at the individual clinician level.
- The numerator includes final reports classified as 'probably benign". 'Probably benign' includes the following classifications: Mammography Quality Standards Act (MQSA) assessment category of "probably benign"; Breast Imaging-Reporting and Data System (BI-RADS®) category 3; or Food and Drug Administration (FDA)-approved equivalent assessment category.
- The denominator includes all final reports for screening mammograms.
- There are no denominator exclusions for this measure.
- ICD-10 and CPT or HCPCS codes included.
- A calculation algorithm describes the process of calculating the performance rate of the measure.
- The developer encourages the results of the measure to be stratified by race, ethnicity, sex, and payer.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*
- *Is the logic or calculation algorithm clear?*

2a2. Reliability Testing Testing attachment

Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

Inter-rater reliability for one year (1/1/2010-12/31/2010) in 3 radiology practice sites and 114 patient records showed 100% agreement for numerator and denominator reliability and overall reliability. Kappa statistics were not calculated because of complete agreement.

Describe any updates to testing: see below

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method of reliability testing:

- The dataset used included Medicare Part B claims data from 2012 – 2014. The number of physicians were 45,558. Of these physicians, 2,750 were from registry. The number of patients reported were 9,705,757.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

Results of reliability testing:

- Measure score reliability results:

Summary of PQRS Reliability Score Stats by Year (2012 - 2014)

Year	Number of Providers	Reliability p25	Reliability median	Reliability p75	Reliability mean	Reliability LCLM	Reliability UCLM
2012	18178	1	1	1	0.99775	0.99746	0.99804
2013	16843	1	1	1	0.99562	0.99515	0.99608
2014	10537	1	1	1	0.99999	0.99999	0.99999
All	45558	1	1	1	0.99748	0.99727	0.99769

- The developer states that the mean (CI) reliability of 0.99748 (0.99727, 0.99769) demonstrated very good reliability.

Guidance from the Reliability Algorithm : Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Appropriate method used (Box 5) →High reliability statistic and scope (Box 6a) →High

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do you agree that the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

- The guidelines do not focus specifically focus on classification of "probably benign".

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing – see updated face validity

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

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

Validity testing method:

- Face validity was assessed using a panel of experts with representation from the ACR Commission on Breast Imaging and the National Mammography Database.

Validity testing results:

- The respondents either agreed or strongly agreed that physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on this measure. *[NQF requires that face validity testing results indicate that the measure as specified can be used to distinguish good from poor quality.]*

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions in the measure.

Questions for the Committee:

- *Do you agree that no exclusions are needed?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer did not provide a statistical analysis to identify statistically significant and meaningful differences in performance measure scores across physicians.

Question for the Committee:

- *Given the data provided in 1b.,does this measure identify meaningful differences about quality across physicians?*

2b6. Comparability of data sources/methods:

- This measure has one set of specifications.

2b7. Missing Data

- The developer did not provide an analysis of missing data to demonstrate that the performance results are not biased or describe an approach for handling missing data to minimize bias.

Guidance from the Validity Algorithm: Specifications somewhat consistent with evidence (Box 1)→Meaningful differences and missing data (threats to validity) not addressed (Box 2) → INSUFFICIENT

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

Rationale: Threats to validity (meaningful differences and missing data) that are applicable to this measure were not addressed

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Reliability is very high. No concerns.

No. This measurement method is valid.

2a2. Reliability Testing

Comments:

Again, reliability is very high.

2b2. Validity Testing

Comments:

Method is valid.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

No.

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.

- This measure is based on clinical registry data and all data elements are available in electronic sources..
- There are no fees to use the measure.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

All are routinely collected. No concerns- very straight forward.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:

- This measure has been included CMS' Physician Quality Reporting System (PQRS) since 2009 and Value Based Payment Modifier.
- The developer lists public reporting and quality improvement with benchmarking as planned use and states that this measure complements another mammography measure on Hospital Compare but does not provide a credible plan for implementation in these programs with expected timeframes.
- This measure has been endorsed since 2008 - per NQF criteria, performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results:

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation:

- The developer is not aware of unintended consequences related to this measure.

***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

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
Not publicly reported, but this is a known entity and highly publicized in the breast imaging community.

Criterion 5: Related and Competing Measures

Related or competing measures

- This measure is related to the Mammography Follow-up Rates (OP-9) - Centers for Medicare and Medicaid Services

Harmonization

- The Mammography Follow-up Rates (OP-9) measure. The period of data collection for OP-9 is only 45 days, and most code 3 recall is 90 or 180 days.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0508

Measure Title: Diagnostic Imaging: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms

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: 3/31/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

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

Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

- ☐ Health 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

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

☒ Process: Using the "probably benign" screening mammography assessment category appropriately

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

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

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

The Inappropriate Use of Probably Benign Code assesses the frequency of using the mammography assessment category of "probably benign" (BIRADS TM 3) when interpreting screening mammograms. This assessment category is reserved for findings that have a <2% likelihood of malignancy but greater than essentially 0% likelihood of malignancy. Results of several studies have emphasized the recommendation not to use the "probably benign" category in interpreting a screening mammography examination. The practice of rendering category 3 assessments directly from screening examination also has been shown to result in adverse health outcomes: 1) unnecessary follow-up of many lesions that could have been promptly assessed as benign, and 2) delayed diagnosis of a small number of cancers that otherwise may have been smaller in size and less likely to be advanced in stage.

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

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

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

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

☒ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

American College of Radiology. ACR practice guideline for the performance of screening and diagnostic mammography. http://www.acr.org/~/media/ACR/Documents/PGTS/guidelines/Screening_Mammography.pdf Revised 2014. Accessed March 10, 2016.

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

Overall final assessment of findings should be based on all imaging studies performed up to that day. In addition, they must be classified according to the FDA-approved final assessment categories [8] and should follow the categories defined in the ACR BI-RADS® 5th edition, 2012 [12] (or any subsequent revisions). The BI-RADS® provides a framework for reporting, lesion assessment, imaging-pathologic correlation, quality improvement, and medical outcomes auditing. (page 6)

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

Not graded.

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

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

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

☐ Yes → ***complete section 1a.7***

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

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

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

Nelson HD, Cantor A, Humphrey L, et al. Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Jan. (Evidence Syntheses, No. 124.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK343819/

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

The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.

These recommendations apply to asymptomatic women aged 40 years or older who do not have preexisting breast cancer or a previously diagnosed high-risk breast lesion and who are not at high risk for breast cancer because of a known underlying genetic mutation (such as a *BRCA1* or *BRCA2* gene mutation or other familial breast cancer syndrome) or a history of chest radiation at a young age.

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

The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.

This recommendation is assigned a Grade B. The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. Suggestions for the practice are to offer or provide this service.

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

GRADE C – The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small.

The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years.

- For women who are at average risk for breast cancer, most of the benefit of mammography results from biennial screening during ages 50 to 74 years. Of all of the age groups, women aged 60 to 69 years are most

likely to avoid breast cancer death through mammography screening. While screening mammography in women aged 40 to 49 years may reduce the risk for breast cancer death, the number of deaths averted is smaller than that in older women and the number of false-positive results and unnecessary biopsies is larger. The balance of benefits and harms is likely to improve as women move from their early to late 40s.

- In addition to false-positive results and unnecessary biopsies, all women undergoing regular screening mammography are at risk for the diagnosis and treatment of noninvasive and invasive breast cancer that would otherwise not have become a threat to their health, or even apparent, during their lifetime (known as "overdiagnosis"). Beginning mammography screening at a younger age and screening more frequently may increase the risk for overdiagnosis and subsequent overtreatment.

- Women with a parent, sibling, or child with breast cancer are at higher risk for breast cancer and thus may benefit more than average-risk women from beginning screening in their 40s.

Go to the Clinical Considerations section for information on implementation of the C recommendation.

Grade I – The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.

The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of screening mammography in women aged 75 years or older.

The USPSTF concludes that the current evidence is insufficient to assess the benefits and harms of digital breast tomosynthesis (DBT) as a primary screening method for breast cancer.

The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of adjunctive screening for breast cancer using breast ultrasonography, magnetic resonance imaging, DBT, or other methods in women identified to have dense breasts on an otherwise negative screening mammogram.

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

Complete section 1a.7

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

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

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

Complete section 1a.7

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

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

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

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

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

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

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

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

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

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

1a.8 OTHER SOURCE OF EVIDENCE

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

Evidence for the measure was based on the ACR Breast Imaging Reporting and Data System (BI-RADS®) Atlas, a quality assurance tool used to standardize reporting, bring clarity to breast imaging interpretations and management recommendations, and to facilitate outcome monitoring.

The BI-RADS Atlas has been in use since 1992 and revised numerous times, with the last update in 2013. It is a dynamic tool that is for practical use in a breast imaging practice with the purpose to allow unambiguous breast imaging reports and meaningful tools to audit and evaluate practice.

Individual studies were also used as supporting evidence as listed in 1a.8.2.

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

A Pubmed search was conducted using the following key words:

"mammography" and "probably benign"

"mammography" and "BI-RADS 3"

"mammography" and BIRADS 3"

"mammography" and "category 3"

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

1. **Finding early invasive breast cancers: a practical approach** (2008):

Harvey JA, Nicholson BT, Cohen MA. Finding early invasive breast cancers: a practical approach. *Radiology.* 2008; 248: 61-76.

http://pubs.rsna.org/doi/10.1148/radiol.2481060339?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed

"The BI-RADS 3 (probably benign) category is a frequently misapplied category (28). Probably benign does not equate to "I am not sure what to do with this lesion." If you are not sure, get more information by obtaining more mammographic views or US images or get input from someone else. The two primary indications for a probably benign assessment are a round or oval mass or grouped round calcifications on a baseline mammogram. It is also okay to consider a lesion probably benign if it is far posterior and that portion of the

breast was not visualized previously. A round or oval mass that is new is not "probably benign" unless it represents a specific benign diagnosis such as a simple cyst or lymph node.

The key to successful application of the BI-RADS 3 category is rigorous evaluation of the lesion characteristics and strict adherence to the criteria differentiating benign from possibly malignant. Many lesions that are assigned a BI-RADS 3 classification and are ultimately determined to be malignant at follow-up examination were, in retrospect, misclassified as BI-RADS 3 initially (29). Findings recalled from screening for diagnostic evaluation should be assigned a category of BI-RADS 0 (needs additional evaluation), not BI-RADS 3, even if the level of suspicion is low that the finding represents a cancer."

2. Does Direct Radiologist-Patient Verbal Communication Affect Follow-Up Compliance of Probably Benign Assessments? (2016): http://www.jacr.org/article/S1546-1440(15)00993-X/pdf

Bosma MS, Neal CH, Klein KA et al. Does Direct Radiologist-Patient Verbal Communication Affect Follow-Up Compliance of Probably Benign Assessments? *JACR*. 2016;13:279-285

"The aim of this study was to determine whether direct verbal communication of results by a radiologist affected follow-up compliance rates for probably benign breast imaging findings. High initial compliance was achieved by radiologist or technologist verbal communication of findings and recommendations. Direct communication by the radiologist did not increase compliance compared with communication by a technologist."

3. Use of BI-RADS 3–Probably Benign Category in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial (2011):

Baum JK, Hanna LG, Suddhasatta A et al, Use of BI-RADS 3–Probably Benign Category in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial. *Radiology*. 2011; 260:1, 61-67.

http://pubs.rsna.org/doi/full/10.1148/radiol.11101285

"Our results confirm the importance of a complete evaluation of lesions before classifying the lesion as probably benign, given the low rate of compliance with this recommendation. When a woman is called back for additional imaging of a lesion, the radiologist has the opportunity to discuss the finding and the importance of short-interval follow-up as that recommendation is given. This approach may also improve compliance with the appropriate follow-up."

4. The probably benign assessment (2007): http://www.ncbi.nlm.nih.gov/pubmed/17888768

Leung JW, Sickles EA, The probably benign assessment. *Radiology Clin North Am.* 2007 45(5): 773-89.

"The probably benign assessment (category 3 in the Breast Imaging Reporting and Data System) is associated with a less than 2% probability of malignancy. Its use in mammography is well supported by robust data from various large-scale prospective studies. Use of the probably benign assessment for lesions visible only at ultrasound or MR imaging is much less well established. This article examines in depth the use of the probably benign assessment: which lesions should be assessed as probably benign, the published evidence supporting such use, pitfalls in misuse, and areas of potentially expanded use that currently are under investigation."

5. Lesion and Patient Characteristics Associated with Malignancy After a Probably Benign Finding on Community Practice Mammography (2008):

Lehman CD, Rutter CM, Eby PR et al. Lesion and Patient Characteristics Associated with Malignancy After a Probably Benign Finding on Community Practice Mammography, *American Journal of Roentgenology*. 2008. 190 (2) 511-515.

Read More: http://www.ajronline.org/doi/full/10.2214/AJR.07.2153?src=recsys

http://www.ajronline.org/doi/full/10.2214/AJR.07.2153?src=recsys

"The cancer yield of 8.8% (150/1,711) for lesions categorized as probably benign in this study population is higher than the expected yield of less than 2% and supports the observation that failure to adhere to a strict set of probably benign lesion characteristics may result in a stronger association with malignancy. The converse findings are also interesting: 27 of 129 cases did meet the strict morphologic criteria for a probably benign assessment. This figure translates to a cancer yield of 1.6% (27/1,711) for all instances of probably benign findings in our study and falls squarely within the accepted published value of less than 2%."

6. Malignant lesions initially subjected to short-term mammographic follow-up:
http://onlinelibrary.wiley.com/doi/10.1002/jmri.21123/full

Rosen EL, Baker JA, Soo MS, Malignant lesions initially subjected to short-term mammographic follow-up. *Radiology*. 2002; 223(1) 221-228.

The authors studied whether systematically evaluated criteria were used for probably benign lesions described within that category. Based on 295 cases that had short-term mammographic follow-up with an eventual biopsy recommendation, it was found that short term mammographic follow-up is often recommended outside of diagnostic criteria for probably benign lesions.

"The findings of this study underscore the need for standardized and universally accepted criteria for the BI-RADS probably benign assessment category. Studies have established the efficacy of the probably benign category, but only when strict diagnostic criteria are applied. Radiologists who choose to use this assessment category in their practice should strive to adhere to these established guidelines so that the desired outcomes of the probably benign category are realized."

7. Breast cancer yield for screening mammographic examinations with recommendation for short interval follow-up:

Kerlikowske K, Smith-Bindman R, Abraham LA, et al. Breast. *Radiology.* 2005;234:684-692. doi:10.1148/radiol.2343031976.

The study compared cancer yield for mammographic screening exams that had short-interval follow-up recommendations after a diagnostic examination versus with only screening examinations. Over 1 million screening examinations over a four year period were collected and reviewed. The authors concluded that many initial screening examinations include short-interval follow-up recommendations based on screening and no diagnostic examination. The cancer yield for these exams is low and lower than with a diagnostic work-up prior to short-interval follow-up, which may result in periodic surveillance of a high number of benign findings.

7. Use of BI-RADS 3—probably benign category in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial:

Baum JK, Hanna LG, Acharyya S, et al. Use of BI-RADS 3—probably benign category in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial. *Radiology.* 2011;260(1):61-67.

This study was conducted to determine the frequency of use of the BI-RADS 3, probably benign category for during the American College of Radiology Imaging Network (ACRIN) Digital Mammographic Imaging Screening Trial (DMIST), either at the time of screening mammography or after work-up; the frequency of patients returning for recommended follow-up; and the rate and stages of any malignancies subsequently found in patients for whom short-term interval follow-up was recommended.

During the DMIST trial, radiologists used the BI-RADS 3 classification at a low rate (2.3% of patients) and tumors assigned a BI-RADS 3 category had a low rate of malignancy and 71% returned for the follow-up. The relatively high rate of noncompliance with short-interval follow-up recommendations supports the practice of radiologists thoroughly evaluate lesions before categorizing as BI-RADS3.

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0508_Evidence_MSF5.0_Data.doc,evidence_attachment_0508-635950377869382975.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
The “probably benign” assessment category is reserved for findings that have
a high probability (=98%) chance of being benign and should not be used as a
category for indeterminate findings. Inappropriate designation of findings as “probably benign” can result in unnecessary follow-up of lesions that could
have been quickly classified or delayed diagnosis and treatment of cancerous lesions. Published guidance documents emphasize the need to conduct a complete diagnostic imaging evaluation before making a probably benign (Category 3 assessment; making it inadvisable to use the probably benign categorization when interpreting a screening mammogram. Immediate completion of a diagnostic imaging evaluation for abnormal screening mammograms eliminates potential anxiety that women would endure with the short interval follow-up that is recommended for “probably benign” findings.The “probably benign” assessment category is reserved for findings that have
a high probability (=98%) chance of being benign and should not be used as a
category for indeterminate findings. Inappropriate designation of findings as “probably benign” can result in unnecessary follow-up of lesions that could
have been quickly classified or delayed diagnosis and treatment of cancerous lesions. Published guidance documents emphasize the need to conduct a complete diagnostic imaging evaluation before making a probably benign (Category 3 assessment; making it inadvisable to use the probably benign categorization when interpreting a screening mammogram. Immediate completion of a diagnostic imaging evaluation for abnormal screening mammograms eliminates potential anxiety that women would endure with the short interval follow-up that is recommended for “probably benign” findings.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

CMS Physician Quality Reporting System

This measure was included in the CMS Physician Quality Reporting System as measure ##146: Radiology: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms from 2009 until now. There is a gap in care as shown by this data; between 2012 and 2014 97.25 % of patients reported on did not meet the measure.

Scores on this measure for 2012-2014 (calculated using data from CMS):
N=45,558 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 97.25% of physicians did not meet the measure (100% - 2.74% who met the measure). Across physicians with at least 10 patients and a performance rate greater than zero for the 3-year period 2012-2014, performance rate= 2.74%.

The performance rate quartiles for the same period 2012-2014 for physicians with at least 10 patients and performance rate >0 were as follows:

Scores on this measure is N= 45,558
25th percentile: 0.33%
50th percentile: 0.74%
75th percentile: 1.96%

Exception Rate: This measure is not specified with exceptions. See attached performance data.

• The performance rate for the three year period was calculated as the count of reported instances where performance was met (numerator=266, 192) divided by the total number of reported instances (9705757). Performance rate was also calculated in this way for each year (2012-2014) with the following results:

2012-2014	2.74%
2012	2.09%
2013	5.48%
2014	0.49%

Additionally, the percentage of eligible professionals who could have reported the measure has remained low until 2014; reporting rate increased from 51.78% in 2012 to 91.73% in 2014. While increased reporting rate increases the potential for patients receiving optimal care, at the 2014 rates there were still 305,702 patients who were not reported on for the measure. Rates below are from PQRS participation data received from CMS for years 2012-2014.

Year	Reporting Rate
2012-2014	66.77%
2012	51.78%
2013	64.95%
2014	91.73%

Rates of success on this measure appear low (as an inverse measure), but the relevant point is the impact on the Hospital Outpatient Quality Reporting (HOQR) program Mammography Follow-up Rates measure (OP-9). Based on claims data the use of code 3 on screening mammograms is 0.49% which is a substantial proportion of the median recall rate of approximately 8% in hospital compare. Variation in compliance with this metric may therefore be expected to result in substantial mis-classification of practices with respect to OP-9. In addition the registry rate of code 3 on screening mammography is 1.56% which is a large enough proportion of the minimally acceptable maximum recall rate standard of 12%. Thus, sites might appear to meet acceptable practice standards, but in fact might be failing to meet minimally acceptable standards by misusing the BIRADS code 3 category. Because the period of data collection for the Hospital compare recall metric is only 45 days, and most code 3 recall is 90 or 180 days, this metric complements the Hospital Compare metric and is essential for its integrity

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.
There is sufficient performance data.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
After a search of the medical literature, no disparities have been identified in the area of inappropriate use of probably benign.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
The American College of Radiology advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data.

A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables (1). A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity (referred to as granular ethnicity and based on one´s ancestry) and language need (a rating of spoken English language proficiency of less than very well and one´s preferred language of health-related encounters)." (2)

1. National Quality Forum Issue Brief (no. 10) Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NWF, August 2008.

2. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
(1) All the previously cited studies emphasize the need to conduct a complete diagnostic imaging evaluation before making a probably benign (category 3) assessment; hence it is recommended not to render such an assessment in interpreting a screening mammography examination. The practice of rendering category 3 assessments directly from screening examination also has been shown to result in adverse outcomes: 1) unnecessary follow - up of many lesions that could have been promptly assessed as benign, and 2)delayed diagnosis of a small number of cancers that otherwise may have been smaller in size and less likely to be advanced in stage.

(2)The “probably benign” assessment category is reserved for findings that have a high probability (=98%) chance of being benign and should not be used as a category for indeterminate findings. Inappropriate designation of findings as “probably benign” can result in unnecessary follow-up of lesions that could have been quickly classified or delayed diagnosis and treatment of cancerous lesions.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
(1) D´Orsi CJ, Bassett LW, Berg WA, et al. BI-RADS: Mammography, 5th edition in: D´Orsi CJ, Mendelson EB, Ikeda DM, et al: Breast Imaging Reporting and Data System: ACR BI-RADS – Breast Imaging Atlas, Reston, VA, American College of Radiology, 2014

(2) Kerlikowske K, Smith-Bindman R, Abraham LA, et al. Breast cancer yield for screening mammographic examinations with

recommendation for short-interval follow-up. Radiology. 2005;234:684-692. doi:10.1148/radiol.2343031976

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not applicable

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 subcriteria 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):*
Cancer, Cancer : Breast, Cancer : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

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.acr.org/Quality-Safety/Quality-Measurement/Medicare-Value-Based-Programs/PQRS-Sample

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Final reports classified as "probably benign"

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
This measure is to be reported each time a screening mammogram is performed during the reporting period. It is anticipated that clinicians who provide the professional component of diagnostic imaging studies for screening mammograms will submit this measure.

S.6. **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, 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.
Numerator Definition:

Probably Benign Classification – Mammography Quality Standards Act (MQSA) assessment category of "probably benign"; Breast Imaging-Reporting and Data System (BI-RADS®) category 3; or Food and Drug Administration (FDA)-approved equivalent assessment category

Numerator Instructions: For performance, a lower percentage, with a definitional target approaching 0%, indicates appropriate assessment of screening mammograms (eg, the proportion of screening mammograms that are classified as "probably benign").

FOR EHR SPECIFICATIONS:
No Current HQMF eCQM Available.

FOR ADMINISTRATIVE CLAIMS SPECIFICATIONS:
Report CPT Category II code: 3343F: Mammogram assessment category of "probably benign", documented

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All final reports for screening mammograms

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
FOR EHR SPECIFICATIONS:
No Current HQMF eCQM Available.

FOR ADMINISTRATIVE CLAIMS SPECIFICATIONS:
Diagnosis for screening mammogram (ICD-9-CM) [for use 1/1/2015-9/30/2015]: V76.11, V76.12
Diagnosis for screening mammogram (ICD-10-CM) [for use 10/01/2015-12/31/2015]: Z12.31
AND
Patient encounter during the reporting period (CPT or HCPCS): 77057, G0202

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No Denominator Exclusions or Denominator Exceptions

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
None

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
We encourage the results of this measure to be stratified by race, ethnicity, sex, and payer.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1) Find the patients who meet the initial patient population (ie, the general group of patients that the performance measure is designed to address).
2) From the patients within the initial patient population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial patient population and denominator are identical.
3) From the patients within the denominator, find the patients who qualify for the Numerator (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator

If the patient does not meet the numerator, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not applicable

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not applicable

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Individual

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Imaging Facility
If other:

S.28. **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.)*
This measure is not included in a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_attachment_146-635936387465194333.docx,NQF146_17_Mar.xlsx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0508

Measure Title: Diagnostic Imaging: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms

Date of Submission: 3/1/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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:	Measure Tested with Data From:

(*must be consistent with data sources entered in S.23*)	
☐ 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*).
We used Medicare Part B administrative claims data for 2012-2014 for the reliability testing.

1.3. What are the dates of the data used in testing? 2012 -2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> 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: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> 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 numbers of physicians were 45,558 physicians
Among these physicians 42,808 were claims and 2,750 were from registry

- The data collection period was 2012-2014
- Data abstraction was performed in 2015

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

- Number of patients eligible were 14,536,696 (avg. per NPI is 319.08)
- Number of patients reported were 9,705,757 (avg. per NPI is 213.04)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Data was only used for reliability testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data

are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

There are no SDS variables for this measure.

_____

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*)

An assessment of measure reliability applying a reliability coefficient in the form of the signal to noise ratio (SNR). In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers.[1] The signal is the variability in measured performance that can be explained by real differences in physician performance and the noise is the total variability in measured performance. Reliability is then the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in physician performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

The SNR reliability testing is performed using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

To estimate these parameters (Alpha and Beta) three steps were used:

1) Build a data file of the proper form for physician-to-physician variance estimation.

2) Use the Betabin SAS macro to estimate the physician-to-physician variance.[2]

3) Use the physician-to-physician variance estimate and the physician-specific information to calculate the physician specific reliability scores.

Reliability can be estimated at different points. The PCPI testing followed the convention of estimating reliability at two points: 1) at a minimum number of qualities reporting events per physician and 2) at the average number of quality reporting events per physician. We generally set the minimum number required as 10 events. Limiting the reliability analysis to only those physicians with a minimum number of events reduces the bias introduced by the inclusion of physicians without a significant numbers of events.

A physician level registry or claims database was used for extracting the relevant physician level information. Conditional on having measure data elements from a large and robust sample of physicians, a deidentified measure reliability analysis can be performed.

1. Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on December 24, 2015.)
2. Wakelin I: MACRO Betabin. [http://www.sensory.org/library/files/SAS/betabin-v22.sas]

Data analysis included these fields:

Reporting Method	N (# of NPIs)	# Patients Eligible	Average Eligible Patients per NPI	# of Patients Reported	Average Patients reported per NPI	# Measure Met	% Measure Met (Mean)	# Exclusions	% Exclusions (Mean)

measure met > 0%								measure NOT met > 0%					
n	Min	p25	Median	p75	Max	# Measure NOT Met	% Measure NOT Met (Mean)	n	Min	p25	Median	p75	Max

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*)

Physician to Physician variation stats- 146

Label	*Estimate*	*StandardError*	*tValue*	*Probt*	*Alpha*	*Lower*	*Upper*
mu	0.03548	0.0006	59.61	<.0001	0.05	0.03431	0.0367
alpha	0.03769	0.00045	84.04	<.0001	0.05	0.03681	0.0386
beta	1.0246	0.01999	51.25	<.0001	0.05	0.9854	1.0637

Summary of PQRS Reliability Score Stats by Year (2012 - 2014)

Year	*Number of Providers*	*Reliability p25*	*Reliability median*	*Reliability p75*	*Reliability mean*	*Reliability LCLM*	*Reliability UCLM*
2012	18178	1	1	1	0.99775	0.99746	0.99804
2013	16843	1	1	1	0.99562	0.99515	0.99608
2014	10537	1	1	1	0.99999	0.99999	0.99999
All	45558	1	1	1	0.99748	0.99727	0.99769

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 mean (CI), P25, median, P75 of the reliability score results are shown in the above table for all 3 years as well as by each year. Our mean (CI) reliability is 0.99748 (0.99727, 0.99769). A reliability of 0.80 is considered very good reliability. So according to the reliability testing analysis, the results demonstrated very good reliability.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

An expert panel was used to assess face validity of the measure. This panel consisted of 20 members, with representation from the ACR Commission on Breast Imaging and the National Mammography Database. The panel was asked to rate their agreement with the following statements:

1. The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.

2. Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.

3. In your opinion, how might this measure contribute to quality improvement? Check all that apply.
   - No value
   - Increase awareness of the appropriate use of mammographic assessment categories for screening mammography exams
   - Is complementary to the recall rate metric used in Hospital Compare, with a 45-day period examined for recall
   - Promotes higher quality management and treatment

ACR Commission on Breast Imaging
Alson, Mark MD
Appleton, Catherine MD
Baker, Jay MD
Hendrick, R. Edward PhD
Lee, Carol MD
Monticciolo, Debra MD
Newell, Mary MD
Parkinson, Brett MD
Rebner, Murray MD
Sickles, Edward MD
Smetherman, Dana MD
Smith, Robert PhD
Warren, Linda MD

ACR National Mammography Database Committee
Rosenberg, Robert MD
Sickles, Edward MD
Berg, Wendie MD
Ellis, Richard MD
Zuley, Margarita MD
Burnside, Elizabeth MD
Patel, Bhavika MD
Lee, Cindy MD

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

The scores obtained from the measure as specified will accurately differentiate quality across providers.

Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.							
Answer Options	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Rating Average	Response Count
	0	1	1	5	4	4.09	11
						answered question	11
						skipped question	0

Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.							
Answer Options	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Rating Average	Response Count
	0	0	0	5	6	4.55	11
						answered question	11
						skipped question	0

In your opinion, how might this measure contribute to quality improvement? Check all that apply.		
Answer Options	Response Percent	Response Count
No value	0.0%	0
Increase awareness of the appropriate use of mammographic assessment categories for screening mammography exams	100.0%	11
Is complementary to the recall rate metric used in Hospital Compare, with a 45-day period examined for recall	54.5%	6
Promotes higher quality management and treatment	81.8%	9
	answered question	11
	skipped question	0

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

The expert panel agreed that the measure remained valid based on existing and new evidence.

1. **The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.**

Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 11 responses, the mean score was 4.09 which places the mean agreement between Agree and Strongly Agree. Only one respondent disagreed and no respondents strongly disagreed.

2. **Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.**

Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 11 responses, the mean score was 4.55 which places the mean agreement between Agree and Strongly Agree. No respondents were neutral and none disagreed or strongly disagreed.

3. **In your opinion, how might this measure contribute to quality improvement? Check all that apply.**

Respondents to this question were able to choose any number of responses. Out of 11 respondents, 100% agreed that this measure would increase awareness of appropriate use, 54.5% believed it was complementary to the recall rate metric in Hospital Compare, and 81.8% believed it would promote higher quality management and treatment. No respondents felt that the measure had no value.

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

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

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

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

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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)

N/A

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

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

N/A

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

N/A
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A

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

N/A

2b4.9. Results of Risk Stratification Analysis:
N/A

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

N/A

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

_____
2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in*

1b)

N/A

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

N/A

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

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

Not applicable.

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

Not applicable.

2b6.3. What is your interpretation of the results in terms of 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*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Not applicable.

2b7.2. 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.

2b7.3. 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.**

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic health records (EHRs)

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.**

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.**

No feasibility assessment **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 ope rational 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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Clarifying definitions and instructions were added to the numerator based on feedback from the PQRS program. This measure was found to be reliable and feasible for implementation.

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

Not applicable

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	Payment Program Physician Quality Reporting System https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html?redirect=/pqri/ Value Based Payment Modifier https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html

4a.1. For each CURRENT use, checked above, provide:

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

This measure has been included in the Physician Quality Reporting System since 2009 as Measure #146. Shown below are national average performance rates as reported in the CMS Report: 2013 Reporting Experience Including Trends (2007-2014) Physician Quality Reporting System and Electronic Prescribing (eRx) Incentive Program, APPENDIX, Table A27. Reporting and Performance Information by Individual Measure for the Physician Quality Reporting System (2010 to 2013).

Year	Average Performance Rate
2010	1.1%
2011	1.5%
2012	0.9%
2013	0.9%

The performance rate was calculated as the count of reported instances where performance was met (numerator) divided by the total number of reported instances that excluded reported exclusions (i.e., performance denominator).
(link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/AnalysisAndPayment.html

Rates of success on this measure appear low (as an inverse measure), but the relevant point is the impact on the Hospital Outpatient Quality Reporting (HOQR) program Mammography Follow-up Rates measure (OP-9). Variation in compliance with this metric may therefore be expected to result in substantial mis-classification of practices with respect to OP-9. Because the period of data collection for the Hospital compare recall metric is only 45 days, and most code 3 recall is 90 or 180 days, this metric complements the Hospital Compare metric and is essential for its integrity.

The ACR believes that the reporting of participation information is a beneficial first step on a trajectory toward the public reporting of performance results, which is appropriate since the measure has been tested and the reliability of the performance data has been validated. Continued NQF endorsement will facilitate our ongoing progress toward this public reporting objective. Additionally, the CMS Physician Compare website is phasing in quality measures over the next several years. Quality measures are tools that help measure health care processes and outcomes. These data are associated with the ability to provide high-quality health care and physician participation in quality programs such as PQRS and the Value Modifier.

4a.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 is an accountability measure and used in the CMS quality and payment programs.

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Scores on this measure for 2012-2014 (calculated using data from CMS):
N=45,558 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 97.25% of physicians did not meet the measure (100% - 2.74% who met the measure). Across physicians with at least 10 patients and a performance rate greater than zero for the 3-year period 2012-2014, mean performance rate= 2.74%.

Additionally, the percentage of eligible professionals who could have reported the measure has remained low until 2014; reporting rate increased from 51.78% in 2012 to 91.73% in 2014. While increased reporting rate increases the potential for patients receiving optimal care, at the 2014 rates there were still 305,702 patients who were not reported on for this measure. Rates below are from PQRS participation data received from CMS for years 2012-2014.

Rationale for Performance Calculations

- Medicare claims data with information on reporting measure #146 from years 2012-2014 was used for performance calculation and analyses.
- For each year, if the patient's eligible (pts_eligible) for a particular physician (npi) was greater or equal to 10, the physician was included in the analysis. For measure 146, among 53,807 physicians 45,558 had at least 10 eligible patients for all 3 years.
- Among 45,558 total physicians included in the analysis, 42,808 submitted data by claims, and 2,750 submitted data by registry (reporting_method). For our analyses we used the combined total of 45,558 for both claims and registry reported cases.

4b.2. **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 is significant improvement from 2012 to 2014 for this measure.
Rates of success on this measure appear low (as an inverse measure), but the relevant point is the impact on the Hospital Outpatient Quality Reporting (HOQR) program Mammography Follow-up Rates measure (OP-9). Based on claims data the use of code 3 on screening mammograms is 0.49% which is a substantial proportion of the median recall rate of approximately 8% in hospital compare. Variation in compliance with this metric may therefore be expected to result in substantial mis-classification of practices with respect to OP-9. In addition the registry rate of code 3 on screening mammography is 1.56% which is a large enough proportion of the minimally acceptable maximum recall rate standard of 12%. Thus, sites might appear to meet acceptable practice standards, but in fact might be failing to meet minimally acceptable standards by misusing the BIRADS code 3 category. Because the period of data collection for the Hospital compare recall metric is only 45 days, and most code 3 recall is 90 or 180 days, this metric complements the Hospital Compare metric and is essential for its integrity.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences related to this measurement.

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)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Mammography Follow-up Rates (OP-9)
Centers for Medicare and Medicaid Services

5a. Harmonization

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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The OP-9 measure is calculated using administrative claims data. The period of data collection for OP-9 is only 45 days, and most code 3 recall is 90 or 180 days.

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.)
There are no competing measures (conceptually both the same measure focus and 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.
Attachment **Attachment:** Related_measure_OP-09_Mamm_TechSpec_042414_-2-.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American College of Radiology (ACR)
Co.2 **Point of Contact:** Judy, Burleson, jburleson@acr.org, 703-648-3787-
Co.3 **Measure Developer if different from Measure Steward:** American College of Radiology

Co.4 **Point of Contact:** Alicia, Blakey, ablakey@acr.org, 703-390-9842-

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.
List of Work Group Members:
William Golden, MD (Co-Chair) (internal medicine)
David Seidenwurm (Co-chair) (diagnostic radiology)
Michael Bettmann, MD
Dorothy Bulas, MD (pediatric radiology)
Rubin I. Cohen, MD, FACP, FCCP, FCCM
Richard T. Griffey, MD, MPH (emergency medicine)
Eric J. Hohenwalter, MD (vascular interventional radiology)
Deborah Levine, MD, FACR (radiology/ultrasound)
Mark Morasch, MD (vascular surgery)
Paul Nagy, MD, PhD (radiology)
Mark R. Needham, MD, MBA (family medicine)
Hoang D. Nguyen (diagnostic radiology/payer representative)
Charles J. Prestigiacomo, MD, FACS (neurosurgery)
William G. Preston, MD, FAAN (neurology)
Robert Pyatt, Jr., MD (diagnostic radiology)
Robert Rosenberg, MD (diagnostic radiology)
David A. Rubin, MD (diagnostic radiology)
B Winfred (B.W.) Ruffner, MD, FACP (medical oncology)
Frank Rybicki, MD, PhD, FAHA (diagnostic radiology)
Cheryl A. Sadow, MD (radiology)
John Schneider, MD, PhD (internal medicine)
Gary Schultz, DC, DACR (chiropractic)
Paul R. Sierzenski, MD, RDMS (emergency medicine)
Michael Wasylik, MD (orthopedic surgery)

Diagnostic Imaging Measure Development Work Group Staff
American College of Radiology: Judy Burleson, MHSA; Alicia Blakey, MS
American Medical Association-convened Physician Consortium for Performance Improvement: Mark Antman, DDS, MBA; Kathleen Blake, MD, MPH; Kendra Hanley, MS; Toni Kaye, MPH; Marjorie Rallins, DPM; Kimberly Smuk, RHIA; Samantha Tierney, MPH; Stavros Tsipas, MA
National Committee for Quality Assurance: Mary Barton, MD

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2017
Ad.3 **Month and Year of most recent revision:** 02, 2015
Ad.4 **What is your frequency for review/update of this measure?** These measures will be updated every 3 years.
Ad.5 **When is the next scheduled review/update for this measure?** 09, 2017

The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] or American College of Radiology (ACR). Neither the AMA, ACR, PCPI, nor its members shall be responsible for any use of the Measures.

The AMA's, PCPI's and National Committee for Quality Assurance's significant past efforts and contributions to the development and updating of the Measures is acknowledged. ACR is solely responsible for the review and enhancement ("Maintenance") of the Measures as of December 31, 2014.
ACR encourages use of the Measures by other health care professionals, where appropriate.

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

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. The AMA, ACR, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:** See copyright statement above.

Ad.8 **Additional Information/Comments:** Coding/Specifications updates occur annually. The ACR has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of the measures. The process can also be activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.

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 #: 0509
Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms
Measure Steward: American College of Radiology
Brief Description of Measure: Percentage of patients undergoing a screening mammogram whose information is entered into a reminder system with a target due date for the next mammogram
Developer Rationale: Although screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older, recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. The use of patient reminders is associated with an increase in screening mammography.Encouraging the implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals.

Any facility that uses less than annual frequency of screening, greatly increases the importance of attendance at each scheduled screening. Even with annual screening recommendations screening does not always occur biennially. This demonstrates the importance of systematic reminders and active patient outreach.

The purpose of screening is to minimize interval or false negative cancers, as these are failures of the screening process. The 2011 article by Bennett, Sellars and Moss (Ref 1)and an earlier work by Woodman,Threlfall and Boggis (ref 2) examine the effect of interval cancer rates (false negative cancers) by time since screen out to three years in the United Kingdom´s triennial screening program. The Interval cancer rates (false negative cases) increase over time (ref 1,2) and begin to approach incidence rates by the third year (ref 2). Thus screening at greater than 2 year intervals will likely have poor overall outcomes in reducing breast cancer mortality. These papers may also underestimate the rate of interval cancers (ref 1) so the actual rates may be higher.
Efforts to ensure regular screening are therefore necessary to eliminate any screening interval beyond 2 years.

1. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. British Journal of Cancer. 2011. 104: 571-577.
2. Woodman CBJ, Threlfall AG, Boggis CR et al. Is the three year breast screening interval too long? Occurrence of interval cancers in NHS breast screening programme's north western region. BMJ. 1995. 310:224-6

Numerator Statement: Patients whose information is entered into a reminder system with a target due date for the next mammogram
Denominator Statement: All patients undergoing a screening mammogram
Denominator Exclusions: Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Registry
Level of Analysis: Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Oct 28, 2008 **Most Recent Endorsement Date:** Oct 28, 2008

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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 2008

- The evidence for this measure was based on guidelines recommendations from the American College of Radiology (ACR) for the performance of screening and diagnostic mammography and the American College of Radiology (ACR) Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas). The strength of evidence was not ranked.

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 following updates:

- The developer provided a brief statement describing the measure focus and health outcome: Recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. Screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older.
- The developer provided a recommendation from the Community Preventive Services Task Force that recommends the use of client reminders to increase screening for breast and cervical cancers on the basis of strong evidence of effectiveness. **Level of Evidence: Recommended.**
    - The Task Force describes "Recommended" as: The systematic review of available studies provides strong or sufficient evidence that the intervention is effective. The categories of "strong" and "sufficient" evidence reflect the Task Force's degree of confidence that an intervention has beneficial effects. They do not directly relate to the expected magnitude of benefits. The categorization is based on several factors, such as study design, number of studies, and consistency of the effect across studies.
- The developer also provided a recommendation from the USPSTF: Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation:
    - This recommendation includes biennial screening mammography for women within different age groups and risks. **Level of evidence:** Varies depending on the recommendation.
- The developer provided a systematic review (not graded) and a summary of the QQC demonstrating the effectiveness of reminder systems in increasing breast cancer screening by mammography.

Exception to evidence

N/A

Guidance from the Evidence Algorithm : Process/structure measure with systematic review and grading of Task Force recommendation (Box 1) → Summary of the quantity, quality, and consistency (QQC) of the body of evidence (Box 4)

Moderate certainty that the net benefit is substantial (Box 5b) → Moderate

Questions for the Committee:

- *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?*

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.

- The developer provided the following data for current performance for 47,866 physicians participating in the CMS Physician Quality Reporting System:

	2012-2014	2012	2013	2014
Performance rate	85.0%	79.4%	86.0%	87.6%
Range	25th percentile: 91.15% 50th percentile: 100% 75th percentile: 100%	--	--	--
Reporting rate of eligible professionals (PQRS)	52.04%	32.12%	50.28%	88.66%

- For endorsement maintenance, NQF asks for performance scores (current and over time), including mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).

Disparities:

- The developer did not provide disparities data from the measure as specified – this is required for endorsement maintenance.
- The developer states that based on 2010 data from the National Health Interview Survey (NHIS) Asian race, low education status, recent immigrant status, and no regular source of medical care or no medical insurance were factors found to reduce the likelihood for a woman to receive a mammogram.

Questions for the Committee:

- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Clinical Electronic claims and registry data

Specifications:

- The level of analysis is at the individual clinician level.
- The numerator is 'patients whose information is entered into a reminder system with a target due date for the next mammogram'.
- The denominator includes 'all patients undergoing a screening mammogram'.
- The denominator exclusions for this measure include 'Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]'.
- ICD-10 and CPT or HCPCS codes included.
- A calculation algorithm describes the process of calculating the performance rate of the measure.
- The developer encourages the results of the measure to be stratified by race, ethnicity, sex, and payer.
- Developer states there are no significant changes since last endorsement (measure was last endorsed in 2008); however, 'medical reason exceptions' were added to the specifications and age constraint removed in 2014.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability for one year (1/1/2010-12/31/2010) in 3 radiology practice sites and 114 patient records showed 100% agreement for numerator and denominator reliability and overall reliablity. Kappa statistics were not calculated because of complete agreement.

Describe any updates to testing: see below

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method of reliability testing:

- The dataset used included Medicare Part B claims data from 2012 – 2014. The number of physicians were 47,866. Of these physicians, 2,486 were from registry. The number of patients reported were 7,554,604.
- The developers used a beta-binomial model to assess the signal-to-noise ratio. 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 physician from another. This is an appropriate test for measure score reliability. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

Results of reliability testing :

- Measure score reliability results:

Year	Number of Providers	Reliability p25	Reliability median	Reliability p75	Reliability mean	Reliability LCLM	Reliability UCLM
2012	19955	1	1	1	0.87513	0.87134	0.87892
2013	18427	0.81469	1	1	0.85736	0.85371	0.86101
2014	9484	0.98538	0.99698	0.99977	0.98146	0.98057	0.98234
All	47866	0.96641	1	1	0.88936	0.88719	0.89152

- The developer states that the mean (CI) reliability of 0.88936 (0.88719, 0.89152) demonstrated very good reliability.

Guidance from the Reliability Algorithm : Precise specifications (Box 1)→Empirical reliability testing (Box 2) →Computed performance scores for measure entities (Box 4) →Appropriate method used (Box 5) →High reliability statistic and scope (Box 6a) → High

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do you agree that the results demonstrate sufficient reliability so that differences in performance can be identified?*

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score as an indicator of quality was systematically assessed by an expert panel. The expert panel agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Describe any updates to validity testing – see updated face validity

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**
☐ **Empirical validity testing of the measure score**

Validity testing method:

- Face validity was assessed using a panel of experts with representation from the ACR Commission on Breast

Imaging and the National Mammography Database.

Validity testing results:

- The respondents generally agreed that physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on this measure. *[NQF requires that face validity testing results indicate that the measure as specified can be used to distinguish good from poor quality.]*

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The exclusion 'medical reason documentation' was added to the specifications in 2014. The developer did not provide an analysis to determine whether the addition of this exclusion affects overall performance scores, the overall number and percentage of individuals excluded, the frequency distribution of exclusions across measured entitities and interpretation of the results demonstrating that this exclusion is needed.

Questions for the Committee:

- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Without a statistical analysis of the exclusions, does the Committee agree that the exclusions/exceptions are of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer did not provide a statistical analysis to identify statistically significant and meaningful differences in performance measure scores across physicians.

Question for the Committee:

- *Given the data provided in 1b.,does this measure identify meaningful differences about quality across physicians?*

2b6. Comparability of data sources/methods:

- This measure has one set of specifications.

2b7. Missing Data

- The developer did not provide an analysis of missing data to demonstrate that the performance results are not biased or describe an approach for handling missing data to minimize bias.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Exclusions, meaningful differences and missing data (threats to validity) not addressed (Box 2) → INSUFFICIENT

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

Rationale: Threats to validity (exclusions, meaningful differences and missing data) that are applicable to this measure were not addressed

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

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.

- This measure is based on clinical registry data and all data elements are available in electronic sources.
- There are no fees to use the measure.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:

- This measure has been included CMS' Physician Quality Reporting System (PQRS) since 2009 and Value Based Payment Modifier. It is also used for quality improvement with benchmarking in the ACR NRDR Qualified Clinical Data Registry.
- The developer lists public reporting and quality improvement with benchmarking as planned use but does not provide a credible plan for implementation in these programs with expected timeframes.
- This measure has been NQF endorsed since 2008. NQF criteria for usability and use is looking for "performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).

Improvement results:

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Unexpected findings (positive or negative) during implementation:

- The developer is not aware of unintended consequences related to this measure.

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

Criterion 5: Related and Competing Measures

Related or competing measures

- Related measure: 2372 : Breast Cancer Screening

Harmonization

- The developer states that the measures are completely harmonized.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0509

Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

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

Date of Submission: 3/31/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

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

Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

- ☐ Health 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

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

☒ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

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

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

The Reminder System for Screening Mammograms supports notifying patients when their next mammogram is due. Recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. Additionally, recent USPSTF lengthening of screening intervals in the USPSTF guideline may affect compliance adversely. The evidence supports the desired health outcome that screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older.

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

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

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

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

☐ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

Community Preventive Services Task Force. Updated recommendations for client- and provider-oriented interventions to increase breast, cervical, and colorectal cancer screening. Am J Prev Med. 2012;43(1):92-96. doi:10.1016/j.ampre.2012.04.008.

http://www.thecommunityguide.org/cancer/screening/client-oriented/reminders.html

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

Task Force Finding

The Community Preventive Services Task Force recommends the use of client reminders to increase screening for breast and cervical cancers on the basis of strong evidence of effectiveness. The Task Force also recommends the use of client reminders to increase colorectal cancer screening with fecal occult blood testing based on strong evidence of effectiveness. Evidence is insufficient, however, to determine effectiveness of client reminders in increasing colorectal cancer screening with other tests (colonoscopy, flexible sigmoidoscopy), because of inconsistent evidence.

http://www.thecommunityguide.org/cancer/screening/client-oriented/RRreminders.html

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

The Community Preventive Services Task Force (Task Force) uses the terms below to describe its findings.

Grade: Recommended

The systematic review of available studies provides strong or sufficient evidence that the intervention is effective.

The categories of "strong" and "sufficient" evidence reflect the Task Force's degree of confidence that an intervention has beneficial effects. They do not directly relate to the expected magnitude of benefits. The categorization is based on several factors, such as study design, number of studies, and consistency of the effect across studies.

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

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

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

☒ Yes → ***complete section 1a.7***

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

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

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

Nelson HD, Cantor A, Humphrey L, et al. Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Jan. (Evidence Syntheses, No. 124.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK343819/

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

Women aged 50 to 74 years

The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.

These recommendations apply to asymptomatic women aged 40 years or older who do not have preexisting breast cancer or a previously diagnosed high-risk breast lesion and who are not at high risk for breast cancer because of a known underlying genetic mutation (such as a *BRCA1* or *BRCA2* gene mutation or other familial breast cancer syndrome) or a history of chest radiation at a young age.

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

The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.

This recommendation is assigned a Grade B. The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. Suggestions for the practice are to offer or provide this service.

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

Women aged 40 to 49 years

GRADE C – The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small.

The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years.

- For women who are at average risk for breast cancer, most of the benefit of mammography results from biennial screening during ages 50 to 74 years. Of all of the age groups, women aged 60 to 69 years are most likely to avoid breast cancer death through mammography screening. While screening mammography in women aged 40 to 49 years may reduce the risk for breast cancer death, the number of deaths averted is smaller than that in older women and the number of false-positive results and unnecessary biopsies is larger. The balance of benefits and harms is likely to improve as women move from their early to late 40s.

- In addition to false-positive results and unnecessary biopsies, all women undergoing regular screening mammography are at risk for the diagnosis and treatment of noninvasive and invasive breast cancer that would otherwise not have become a threat to their health, or even apparent, during their lifetime (known as "overdiagnosis"). Beginning mammography screening at a younger age and screening more frequently may increase the risk for overdiagnosis and subsequent overtreatment.

- Women with a parent, sibling, or child with breast cancer are at higher risk for breast cancer and thus may benefit more than average-risk women from beginning screening in their 40s.

Go to the Clinical Considerations section for information on implementation of the C recommendation.

Grade I – The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.

Women aged 75 years or older

Grade I - The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of screening mammography in women aged 75 years or older.

AI Women - The USPSTF concludes that the current evidence is insufficient to assess the benefits and harms of digital breast tomosynthesis (DBT) as a primary screening method for breast cancer.

Women with dense breasts - The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of adjunctive screening for breast cancer using breast ultrasonography, magnetic resonance imaging, DBT, or other methods in women identified to have dense breasts on an otherwise negative screening mammogram.

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

http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/breast-cancer-screening1

Complete section 1a.7

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

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

Baron RC, Melillo S, Rimer BK, Coates RJ, Kerner J, Habarta N, Chattopadhyay S, Sabatino SA, Elder R, Leeks KJ, Task Force on Community Preventive Services. Intervention to increase recommendation and delivery of screening for breast, cervical, and colorectal cancers by healthcare providers: a systematic review of provider reminders. *Am J Prev Med* 2010;38(1):110-7.

Baron RC, Rimer BK, Breslow RA, et al. Client-directed interventions to increase community demand for breast, cervical, and colorectal cancer screening: a systematic review. *Am J Prev Med* 2008;35(1S):34-55.

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

Systematic review #1 (Baron et al 2010):

http://www.thecommunityguide.org/cancer/screening/provider-oriented/InterventionsIncreaseRecommendationDeliveryScreeningBreastCervicalColorectalCancersHealthcareProvidersSystematicReview_2.pdf

Systematic review #2 (Baron et al 2008)

http://www.thecommunityguide.org/cancer/screening/client-oriented/Cancer2008_ClientDirected_Demand.pdf

Complete section 1a.7

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

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

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

Systematic review #1 (Baron et al 2010)

This report presents results of systematic reviews of effectiveness, applicability, economic efficiency, barriers to implementation, and other harms or benefits of provider reminder/recall interventions to increase screening for breast, cervical, and colorectal cancers. Evidence in this review of studies published from 1986 through 2004 indicates that reminder/recall systems can effectively increase screening with mammography, Pap, fecal occult blood tests, and flexible sigmoidoscopy.

Systematic review #2 (Baron et al 2008)

This report presents the results of systematic reviews of effectiveness, applicability, economic efficiency, barriers to implementation, and other harms or benefits of interventions designed to increase screening for breast, cervical, and colorectal cancers by increasing community demand for these services. Evidence from these reviews indicates that screening for breast cancer (mammography) and cervical cancer (Pap test) has been effectively increased by use of client reminders, small media, and one-on-one education.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade: No grade.

The systematic reviews identified in this application use a similar evaluation of studies for review *Guide to Community Preventive Services*. Each study is characterized based on both the suitability of study design for assessing effectiveness and the quality of study execution. Study designs are classified using a standard algorithm

Greatest - concurrent comparison groups *and* prospective measurement of exposure and outcome

Moderate - all retrospective designs *or* multiple pre or post measurements but no concurrent comparison group

Least - single pre and post measurements and no concurrent comparison group *or* exposure and outcome measured in a single group at the same point in time

Quality of Execution

Each study is categorized as having good, fair, or limited quality of execution based on the number of limitations noted, studies with 0–1, 2–4, and 5 or more limitations are categorized as having good, fair, and limited execution respectively. Studies with limited execution are not included in bodies of evidence to support recommendations. In general, information on quality of study execution is based only on information in published reports because bias could be introduced based on limited availability or variable quality of additional information from the authors and because collecting additional information from the authors may not be feasible.

Several principles guided the designation of bodies of evidence of effectiveness as strong, sufficient, or insufficient evidence. Strong or sufficient evidence can be based either on a small number of studies with better execution and more suitable design or a larger number of studies with less suitable design or weaker execution

Table 2. Assessing the strength of a body of evidence on effectiveness of population-based interventions in the *Guide to Community Preventive Services*

Evidence of effectiveness[a]	Execution—good or fair[b]	Design Suitability—Greatest, moderate, or least	Number of studies	Consistent[c]	Effect size[d]	Expert opinion[e]
Strong	Good	Greatest	At Least 2	Yes	Sufficient	Not Used
	Good	Greatest or Moderate	At Least 5	Yes	Sufficient	Not Used
	Good or Fair	Greatest	At Least 5	Yes	Sufficient	Not Used
	Meet Design, Execution, Number and Consistency Criteria for Sufficient But Not Strong Evidence				Large	Not Used
Sufficient	Good	Greatest	1	Not Applicable	Sufficient	Not Used
	Good or Fair	Greatest or Moderate	At Least 3	Yes	Sufficient	Not Used
	Good or Fair	Greatest, Moderate, or Least	At Least 5	Yes	Sufficient	Not Used
Expert Opinion	Varies	Varies	Varies	Varies	Sufficient	Supports a Recommendation
Insufficient[f]	A. Insufficient Designs or Execution		B. Too Few Studies	C. Inconsistent	D. Small	E. Not Used

[a]The categories are not mutually exclusive; a body of evidence meeting criteria for more than one of these should be categorized in the highest possible category.
[b]Studies with limited execution are not used to assess effectiveness.
[c]Generally consistent in direction and size.
[d]Sufficient and large effect sizes are defined on a case-by-case basis and are based on Task Force opinion.
[e]Expert opinion will not be routinely used in the *Guide* but can affect the classification of a body of evidence as shown.
[f]Reasons for determination that evidence is insufficient will be described as follows: A. Insufficient designs or executions, B. Too few studies, C. Inconsistent. D. Effect size too small, E. Expert opinion not used. These categories are not mutually exclusive and one or more of these will occur when a body of evidence fails to meet the criteria for strong or sufficient evidence.

Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based Guide to Community Preventive Services— methods. Am J Prev Med 2000;18(1S):35– 43. http://www.thecommunityguide.org/about/methods-ajpm-developing-guide.pdf

Zaza S, Wright-De Aguero LK, Briss PA, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Am J Prev Med 2000;

18(1S):44–74

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

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

Date range:

Baron et al 2010 : 1986-2004

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

Systematic review #1 (Baron et al 2010)

The search for evidence identified 38 studies that reported on using provider reminders to increase recommended screening for breast, cervical, and colorectal cancers. Of these, six were excluded because of their low quality of execution and six more were excluded because of the lack of a concurrent comparison group. Of the 26 remaining studies that qualified for review, five had good quality of execution, and 21 studies had fair quality of execution. Of the studies that qualified for review 9 were from randomized control trials and 16 were observational studies.

Systematic review #2 (Baron et al 2008)

The searches for evidence identified 39 studies of greatest design suitability were identified that reported using client reminders to increase breast cancer screening by mammography. Of these, nine studies were excluded due to limited quality of execution and were excluded because comparison groups received different reminders or reminders of lesser intensity than study groups. Of the 19 remaining studies that qualified for review, had fair quality of execution and two had good quality of execution. Six studies, five classified as good and one as very good met inclusion criteria for cost-effectiveness analysis of client reminders in increasing breast cancer screening by mammography.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

2010 Baron et al

The qualifying studies examined mammography Pap, and colorectal screening. All measured outcomes (screening tests completed, or screening tests recommended or ordered but not necessarily completed) were ascertained by record review.13 studies for mammography – pertained to the primary outcome of interest, completed screening tests.

Mammography screening increased by a median of 10.0% (IQI, 3.0%–19.0) for all screening modalities, but in particular for mammography, the absolute effect of provider reminders on completed screenings appears to have diminished over time. Because background screening rates often were not provided for study populations, the role, if any, of temporal changes in baseline screening rates on these results could not be determined. Evidence in this review of studies published from 1986 through 2004 indicates that reminder/recall systems can effectively increase screening with mammography, Pap, fecal occult blood tests, and flexible sigmoidoscopy.

Baron et al 2008

Twenty studies were identified that reported using small media to increase breast cancer screening by mammography. One study was excluded due to limited quality of execution. Of 19 qualifying studies, 17 had greatest design suitability, of which three had good quality of execution and 14 had fair quality of execution. Two qualifying studies, one with moderate and one with least suitable study design, had fair quality of

execution. Five studies evaluated tailored interventions, twelve evaluated untailored interventions, and two studies included both a tailored and an untailored intervention.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

Systematic review #1 (Baron et al 2010)

The original review included 19 studies. This update included an additional 6 studies. Combined evidence from both the original and the updated review showed the following.

• Mammography screening: median increase of 14.0 percentage points (interquartile interval [IQI]: 2.0 to 24.0 percentage points; 19 studies with 32 study arms).

• Recent mammography screening: median increase of 12.3 percentage points (IQI: 3.0 to 18.9 percentage points; 30 study arms).

• Repeat mammography screening: median increase of 6.0 percentage points (IQI 3.0 to 19.1 percentage points; 8 study arms).

• Enhanced and telephone reminders showed a greater increase (15.5 percentage points [IQI 7.0 to 29.0 percentage points]; 20 study arms) than written reminders alone (4.5 percentage points [IQI: 1.9 to 14.0 percentage points]; 14 study arms).

• When added to other types of interventions, the median incremental effect for client reminders was an increase of 5.0 percentage points (IQI 1.6 to 6.7 percentage points; 12 study arms).

Client reminder interventions to increase breast cancer screening should be applicable across a range of settings and populations, provided they are adapted to the target population and delivery context.

Systematic Review #2 (Baron et al 2008)

According to *Community Guide* methods, there is strong evidence that client reminders increase breast and cervical cancer screening by mammography and Pap test, respectively. These findings should apply across a range of settings and populations. Although evidence also suggests that enhancement of simple printed reminders with additional messages or support to clients results in greater effectiveness, particularly for breast cancer screening, it is not yet known whether such enhancement increases effectiveness among women who have never been screened or who may be hard to reach.

Overall, the median post-intervention increase in completed mammography was 14.0 percentage points (interquartile interval [IQI]= 2.0, 24.0). The magnitude of this effect and consistent positive results across studies and reminder systems demonstrate the effectiveness of client reminders in increasing breast cancer screening by mammography.

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

No reports of benefits or harms related to the use of provider reminders were found. Potential benefits include increases in the use of other preventive services linked to the reminder system.

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

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

1a.8 OTHER SOURCE OF EVIDENCE

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

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

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

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

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

0509_Evidence_MSF5.0_Data.doc,evidence_attachment_0509-635950347428110665.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
Although screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older, recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. The use of patient reminders is associated with an increase in screening mammography.Encouraging the implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals.

Any facility that uses less than annual frequency of screening, greatly increases the importance of attendance at each scheduled screening. Even with annual screening recommendations screening does not always occur biennially. This demonstrates the importance of systematic reminders and active patient outreach.

The purpose of screening is to minimize interval or false negative cancers, as these are failures of the screening process. The 2011 article by Bennett, Sellars and Moss (Ref 1)and an earlier work by Woodman,Threlfall and Boggis (ref 2) examine the effect of interval cancer rates (false negative cancers) by time since screen out to three years in the United Kingdom´s triennial screening program. The Interval cancer rates (false negative cases) increase over time (ref 1,2) and begin to approach incidence rates by the third year (ref 2). Thus screening at greater than 2 year intervals will likely have poor overall outcomes in reducing breast cancer mortality. These papers may also underestimate the rate of interval cancers (ref 1) so the actual rates may be higher.
Efforts to ensure regular screening are therefore necessary to eliminate any screening interval beyond 2 years.

1. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. British Journal of Cancer. 2011. 104: 571-577.
2. Woodman CBJ, Threlfall AG, Boggis CR et al. Is the three year breast screening interval too long? Occurrence of interval cancers in NHS breast screening programme's north western region. BMJ. 1995. 310:224-6

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
CMS Physician Quality Reporting System
This measure was included in the CMS Physician Quality Reporting System as measure #225 Reminder System for Screening Mammograms from 2009 until now. There is a gap in care as shown by this data; between 2012 and 2014 15.0 % of patients reported on did not meet the measure.

Scores on this measure for 2012-2014 (calculated using data from CMS):
N=47,866 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 15.0% of physicians did not meet the measure (100% - 85.0% who met the measure). Across physicians with at least 10 patients and a performance rate greater than zero for the 3-year period 2012-2014, mean performance rate= 85.0%.

The performance rate quartiles for the same period 2012-2014 for physicians with at least 10 patients and performance rate >0 were as follows:

Scores on this measure is N= 47,866
25th percentile: 91.15%

50th percentile: 100%
75th percentile: 100%

Exception Rate: This measure is not specified with exceptions. See attached performance data.

• The performance rate for the three year period was calculated as the count of reported instances where performance was met (numerator=6,423,710) divided by the total number of reported instances (7,554,604). Performance rate was also calculated in this way for each year (2012-2014) with the following results:

2012-2014	85.0%
2012	79.4%
2013	86.0%
2014	87.6%

Additionally, the percentage of eligible professionals who could have reported the measure has remained low until 2014; reporting rate increased from 32.12% in 2012 to 88.66% in 2014. While increased reporting rate increases the potential for patients receiving optimal care, at the 2014 rates there were still 379,320 patients who were potentially not receiving optimal care per the measure. Rates below are from PQRS participation data received from CMS for years 2012-2014.

Year	Reporting Rate
2012-2014	52.04%
2012	32.12%
2013	50.28%
2014	88.66%

Rationale for Performance Calculations

- Medicare claims data with information on reporting measure #225 from years 2012-2014 was used for performance calculation and analyses.

- For each year, if the patient's eligible (pts_eligible) for a particular physician (npi) was greater or equal to 10, the physician was included in the analysis. For measure 225, among 57833 physicians 47866 had at least 10 eligible patients for all 3 years.

- Among 47866 total physicians included in the analysis, 45380 submitted data by claims, and 2486 submitted data by registry (reporting_method). For our analyses we used the combined total of 47866 for both claims and registry reported cases.

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.
There is sufficient performance data.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Many American women do not receive mammograms at recommended intervals, as illustrated by 2010 data from the National Health Interview Survey (NHIS) which found that only 72% of women reported receiving a mammogram within the recommended two-year interval. Additional factors found to reduce the likelihood for a woman to receive a mammogram include Asian race, low education status, and recent immigrant status. Low mammography use was also noted for women who reported having no regular source of medical care or having no medical insurance.

Centers for Disease Control and Prevention (CDC). Cancer screening—United States, 2010. MMWR 2012;61(3):41-45. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6103a1.htm. Accessed 2/3/2014.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
The American College of Radiology advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data.

A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables (1). A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity (referred to as granular ethnicity and based on one´s ancestry) and language need (a rating of spoken English language proficiency of less than very well and one´s preferred language of health-related encounters)." (2)

1. National Quality Forum Issue Brief (no. 10) Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NWF, August 2008.

2. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, Frequently performed procedure

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Breast cancer mortality is reduced with mammography screening, although estimates are of borderline statistical significance, the magnitudes of effect are small for younger ages, and results vary depending on how cases were accrued in trials. Higher stage tumors are also reduced with screening for age 50 years and older. False-positive results are common in all age groups, and are higher for younger women and those with risk factors. Approximately 11 to 22 percent of cases may be over diagnosed. Observational studies indicate that tomosynthesis with mammography reduces recalls, but increases biopsies and cancer detection. Mammography screening at any age is a trade off of a continuum of benefits and harms.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Nelson HD, Cantor A, Humphrey L, et al. Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Jan. (Evidence Syntheses, No. 124.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK343819/

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable

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 subcriteria 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):*

Cancer, Cancer : Breast, Cancer : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Care Coordination

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.acr.org/Quality-Safety/Quality-Measurement/Medicare-Value-Based-Programs/PQRS-Sample and
http://www.acr.org/Quality-Safety/Quality-Measurement/Performance-Measures

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
There are no significant changes since last endorsement.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients whose information is entered into a reminder system with a target due date for the next mammogram

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
This measure is to be reported each time a screening mammogram is performed during the reporting period for patients seen during the reporting period. This measure is intended to reflect the quality of services provided for reminding patients when follow-up mammograms are due.

S.6. **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, 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.
Numerator Note: The reminder system should be linked to a process for notifying patients when their next mammogram is due and should include the following elements at a minimum: patient identifier, patient contact information, dates(s) of prior screening mammogram(s) (if known), and the target due date for the next mammogram

FOR ELECTRONIC SPECIFICATIONS:
Not Applicable

FOR ADMINISTRATIVE CLAIMS SPECIFICATIONS:
Report CPT II Code 7025F: Patient information entered into a reminder system with a target due date for the next mammogram

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients undergoing a screening mammogram

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
FOR ELECTRONIC SPECIFICATIONS:

Not Applicable

FOR ADMINISTRATIVE CLAIMS SPECIFICATIONS:
Diagnosis for mammogram screening (ICD-9-CM) [for use 1/1/2015-9/30/2015]: V76.11, V76.12
Diagnosis for mammogram screening (ICD-10-CM) [for use 10/01/2015-12/31/2015]: Z12.31
AND
Patient encounter during the reporting period (CPT or HCPCS): 77057, G0202

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
FOR ELECTRONIC SPECIFICATIONS:
Not Applicable

FOR ADMINISTRATIVE CLAIMS SPECIFICATIONS:
Report CPT II Code 7025F-1P: Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
We encourage the results of this measure to be stratified by race, ethnicity, sex, and payer.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)

S.16. Type of score:
Rate/proportion
If other:

S.17. 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.18. Calculation Algorithm/Measure Logic *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1) Find the patients who meet the initial patient population (ie, the general group of patients that the performance measure is designed to address).

2) From the patients within the initial patient population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial patient population and denominator are identical.
3) From the patients within the denominator, find the patients who qualify for the Numerator (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator

If the patient does not meet the numerator, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample or survey.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
N/A

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Individual

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility, Imaging Facility
If other:

S.28. **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.)*
This is not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_attachment_225-635936386333704574.docx,NQF_225_17_Mar.xlsx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0509

Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

Date of Submission: 3/2/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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*).
We used Medicare Part B administrative claims data for 2012-2014 for the reliability testing.

1.3. What are the dates of the data used in testing? 2012 -2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> 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: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> 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 numbers of physicians were 47,866 physicians
Among these physicians 45,380 were claims and 2,486 were from registry

- The data collection period was 2012-2014
- Data abstraction was performed in 2015

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

- Number of patients eligible were 14,515,814 (avg. per NPI is 303.26)
- Number of patients reported were 7,554,604 (avg. per NPI is 157.83)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Data was only used for reliability testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

There are no SDS variables for this measure.

_____

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*)

An assessment of measure reliability applying a reliability coefficient in the form of the signal to noise ratio (SNR). In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers.[1] The signal is the variability in measured performance that can be explained by real differences in physician performance and the noise is the total variability in measured performance. Reliability is then the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in physician performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

The SNR reliability testing is performed using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

To estimate these parameters (Alpha and Beta) three steps were used:

1) Build a data file of the proper form for physician-to-physician variance estimation.

2) Use the Betabin SAS macro to estimate the physician-to-physician variance.[2]

3) Use the physician-to-physician variance estimate and the physician-specific information to calculate the physician specific reliability scores.

Reliability can be estimated at different points. The PCPI testing followed the convention of estimating reliability at two points: 1) at a minimum number of qualities reporting events per physician and 2) at the average number of quality reporting events per physician. We generally set the minimum number required as 10 events. Limiting the reliability analysis to only those physicians with a minimum number of events reduces the bias introduced by the inclusion of physicians without a significant numbers of events.

A physician level registry or claims database was used for extracting the relevant physician level information. Conditional on having measure data elements from a large and robust sample of physicians, a deidentified measure reliability analysis can be performed.

1. Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on December 24, 2015.)
2. Wakelin I: MACRO Betabin. [http://www.sensory.org/library/files/SAS/betabin-v22.sas]

Data analysis included these fields:

Reporting Method	N (# of NPIs)	# Patients Eligible	Average Eligible Patients per NPI	# of Patients Reported	Average Patients reported per NPI	# Measure Met	% Measure Met (Mean)	# Exclusions	% Exclusions (Mean)

measure met > 0%								measure NOT met > 0%					
n	Min	p25	Median	p75	Max	# Measure NOT Met	% Measure NOT Met (Mean)	n	Min	p25	Median	p75	Max

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*)

Physician to Physician variation stats- 225

Label	*Estimate*	*StandardError*	*tValue*	*Probt*	*Alpha*	*Lower*	*Upper*
mu	0.2867	0.00168	170.77	<.0001	0.05	0.2834	0.2899
alpha	0.0693	0.0006	115.34	<.0001	0.05	0.06813	0.0705
beta	0.1725	0.00147	117.15	<.0001	0.05	0.1696	0.1754

Summary of PQRS Reliability Score Stats by Year (2012 - 2014)

Year	*Number of Providers*	*Reliability p25*	*Reliability median*	*Reliability p75*	*Reliability mean*	*Reliability LCLM*	*Reliability UCLM*
2012	19955	1	1	1	0.87513	0.87134	0.87892
2013	18427	0.81469	1	1	0.85736	0.85371	0.86101
2014	9484	0.98538	0.99698	0.99977	0.98146	0.98057	0.98234
All	47866	0.96641	1	1	0.88936	0.88719	0.89152

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 mean (CI), P25, median, P75 of the reliability score results are shown in the above table for all 3 years as well as by each year. Our mean (CI) reliability is 0.88936 (0.88719, 0.89152). A reliability of 0.80 is considered very good reliability. So according to the reliability testing analysis, the results demonstrated very good reliability.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

An expert panel was used to assess face validity of the measure. This panel consisted of 20 members, with representation from the ACR Commission on Breast Imaging and the National Mammography Database. The panel was asked to rate their agreement with the following statements:

1. The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.

2. Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.

3. In your opinion, how might this measure contribute to quality improvement? Check all that apply.
   - No value
   - Implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals
   - Is complementary to the USPSTF guideline for screening mammograms to reduce breast cancer mortality
   - Promotes higher quality management and treatment

ACR Commission on Breast Imaging

Alson, Mark MD
Appleton, Catherine MD
Baker, Jay MD
Hendrick, R. Edward PhD
Lee, Carol MD
Monticciolo, Debra MD
Newell, Mary MD
Parkinson, Brett MD
Rebner, Murray MD
Sickles, Edward MD
Smetherman, Dana MD
Smith, Robert PhD
Warren, Linda MD
Rosenberg, Robert MD
Sickles, Edward MD
Berg, Wendie MD
Ellis, Richard MD
Zuley, Margarita MD
Burnside, Elizabeth MD
Patel, Bhavika MD
Lee, Cindy

ACR National Mammography Database Committee

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
The scores obtained from the measure as specified will accurately differentiate quality across providers.
Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.							
Answer Options	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**Rating Average**	**Response Count**
	0	0	0	3	7	4.70	10
answered question							**10**
skipped question							**0**

Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.							
Answer Options	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**Rating Average**	**Response Count**
	0	1	2	3	4	4.00	10
answered question							**10**
skipped question							**0**

In your opinion, how might this measure contribute to quality improvement? Check all that apply.		
Answer Options	**Response Percent**	**Response Count**
No value	0.0%	0
Implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals	90.0%	9
Is complementary to the USPSTF guideline for screening mammograms to reduce breast cancer mortality	30.0%	3
Promotes higher quality management and treatment	70.0%	7
answered question		**10**
skipped question		**0**

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

The expert panel agreed that the measure remained valid based on existing and new evidence.

1. **The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.**

   Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 10 responses, the mean score was 4.70 which placed the mean agreement between Agree and Strongly Agree. No responses were neutral and none disagreed or strongly disagreed. Seven of 10 strongly agreed.

2. **Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.**

   Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 10 responses, the mean score was 4.00 which are consistent with general agreement. One respondent disagreed, 2 were neutral, and the remaining 7 agreed or strongly agreed.

3. **In your opinion, how might this measure contribute to quality improvement? Check all that apply.**

   Respondents were asked to check all statements with which they agreed. 90% felt that implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals, 30% felt this measure is complementary to USPSTF guidelines for screening mammograms and 70% believe this measure will promote higher quality management and treatment. No respondents felt this measure had zero value.

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

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

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

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

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

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

2b4.9. Results of Risk Stratification Analysis:

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

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

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

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

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

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

Not applicable.

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

Not applicable.

2b6.3. What is your interpretation of the results in terms of 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*)

Not applicable.

_____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Not applicable.

2b7.2. 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.

2b7.3. 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.**

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic health records (EHRs)

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.**

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Clarifying definitions and instructions were added to the numerator based on feedback from the PQRS program. This measure was found to be reliable and feasible for implementation.

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

Not applicable

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	Payment Program Physician Quality Reporting System https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html?redirect=/pqri/ Value Based Payment Modifier https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) ACR NRDR Qualified Clinical Data Registry www.acr.org/qcdr

4a.1. For each CURRENT use, checked above, provide:

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

This measure has been included in the Physician Quality Reporting System since 2009 as Measure #225. Shown below are national average performance rates as reported in the CMS Report: 2013 Reporting Experience Including Trends (2007-2014) Physician Quality Reporting System and Electronic Prescribing (eRx) Incentive Program, APPENDIX, Table A27. Reporting and Performance Information by Individual Measure for the Physician Quality Reporting System (2010 to 2013).

Year	Average Performance Rate
2010	N/A
2011	68.5 %
2012	74.6 %
2013	81.6%

The performance rate was calculated as the count of reported instances where performance was met (numerator) divided by the total number of reported instances that excluded reported exclusions (i.e., performance denominator).
(link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/AnalysisAndPayment.html

While these rates do show a steady increase in performance, the 2013 score still indicates that 18.4% of patients reported on did not receive optimal care.

The ACR believes that the reporting of participation information is a beneficial first step on a trajectory toward the public reporting of performance results, which is appropriate since the measure has been tested and the reliability of the performance data has been validated. Continued NQF endorsement will facilitate our ongoing progress toward this public reporting objective. Additionally, the CMS Physician Compare website is phasing in quality measures over the next several years. Quality measures are tools that help measure health care processes and outcomes. These data are associated with the ability to provide high-quality health care and

physician participation in quality programs such as PQRS and the Value Modifier.

4a.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 is an accountability measure and used in the CMS quality and payment programs.

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Scores on this measure for 2012-2014 (calculated using data from CMS):
N=47,866 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 15.0% of physicians did not meet the measure (100% - 85.0% who met the measure). Across physicians with at least 10 patients and a performance rate greater than zero for the 3-year period 2012-2014, mean performance rate= 85.0%.

Additionally, the percentage of eligible professionals who could have reported the measure has remained low until 2014; reporting rate increased from 32.12% in 2012 to 88.66% in 2014. While increased reporting rate increases the potential for patients receiving optimal care, at the 2014 rates there were still 379,320 patients who were potentially not receiving optimal care per the measure. Rates below are from PQRS participation data received from CMS for years 2012-2014.

Rationale for Performance Calculations

- Medicare claims data with information on reporting measure #225 from years 2012-2014 was used for performance calculation and analyses.

- For each year, if the patient's eligible (pts_eligible) for a particular physician (npi) was greater or equal to 10, the physician was included in the analysis. For measure 225, among 57833 physicians 47866 had at least 10 eligible patients for all 3 years.

- Among 47866 total physicians included in the analysis, 45380 submitted data by claims, and 2486 submitted data by registry (reporting_method). For our analyses we used the combined total of 47866 for both claims and registry reported cases.

4b.2. **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 is significant improvement from 2012 to 2014 for this measure.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any unintended consequences related to this measurement.

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)**
2372 : Breast Cancer Screening

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

5a. Harmonization

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 completely harmonized?
Yes

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.)
There are no competing measures (conceptually both the same measure focus and 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.
Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American College of Radiology
Co.2 **Point of Contact:** Judy, Burleson, jburleson@acr.org, 703-648-3787-
Co.3 **Measure Developer if different from Measure Steward:** American College of Radiology
Co.4 **Point of Contact:** Alicia, Blakey, ablakey@acr.org, 703-390-9842-

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.

List of Work Group Members:
William Golden, MD (Co-Chair) (internal medicine)
David Seidenwurm (Co-chair) (diagnostic radiology)
Michael Bettmann, MD
Dorothy Bulas, MD (pediatric radiology)
Rubin I. Cohen, MD, FACP, FCCP, FCCM
Richard T. Griffey, MD, MPH (emergency medicine)
Eric J. Hohenwalter, MD (vascular interventional radiology)
Deborah Levine, MD, FACR (radiology/ultrasound)
Mark Morasch, MD (vascular surgery)
Paul Nagy, MD, PhD (radiology)
Mark R. Needham, MD, MBA (family medicine)
Hoang D. Nguyen (diagnostic radiology/payer representative)
Charles J. Prestigiacomo, MD, FACS (neurosurgery)
William G. Preston, MD, FAAN (neurology)
Robert Pyatt, Jr., MD (diagnostic radiology)
Robert Rosenberg, MD (diagnostic radiology)
David A. Rubin, MD (diagnostic radiology)
B Winfred (B.W.) Ruffner, MD, FACP (medical oncology)
Frank Rybicki, MD, PhD, FAHA (diagnostic radiology)
Cheryl A. Sadow, MD (radiology)
John Schneider, MD, PhD (internal medicine)
Gary Schultz, DC, DACR (chiropractic)
Paul R. Sierzenski, MD, RDMS (emergency medicine)
Michael Wasylik, MD (orthopedic surgery)

Diagnostic Imaging Measure Development Work Group Staff
American College of Radiology: Judy Burleson, MHSA; Alicia Blakey, MS
American Medical Association-convened Physician Consortium for Performance Improvement: Mark Antman, DDS, MBA; Kathleen Blake, MD, MPH; Kendra Hanley, MS; Toni Kaye, MPH; Marjorie Rallins, DPM; Kimberly Smuk, RHIA; Samantha Tierney, MPH; Stavros Tsipas, MA
National Committee for Quality Assurance: Mary Barton, MD

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 02, 2015
Ad.4 **What is your frequency for review/update of this measure?** These measures will be updated every 3 years.
Ad.5 **When is the next scheduled review/update for this measure?** 09, 2017

The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] or American College of Radiology (ACR). Neither the AMA, ACR, PCPI, nor its members shall be responsible for any use of the Measures.

The AMA's, PCPI's and National Committee for Quality Assurance's significant past efforts and contributions to the development and updating of the Measures is acknowledged. ACR is solely responsible for the review and enhancement ("Maintenance") of the Measures as of December 31, 2014.
ACR encourages use of the Measures by other health care professionals, where appropriate.

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

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. The AMA, ACR, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:** See copyright statement above.

Ad.8 **Additional Information/Comments:** Coding/Specifications updates occur annually. The ACR has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of the measures. The process can also be activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.

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 #: 0559

Measure Title: Combination chemotherapy is recommended or administered within 4 months (120 days) of diagnosis for women under 70 with AJCC T1cN0M0, or Stage IB - III hormone receptor negative breast cancer.

Measure Steward: Commission on Cancer, American College of Surgeons

Brief Description of Measure: Percentage of female patients, age >18 at diagnosis, who have their first diagnosis of breast cancer (epithelial malignancy), at AJCC stage T1cN0M0 (tumor greater than 1 cm), or Stage IB -III, whose primary tumor is progesterone and estrogen receptor negative recommended for multiagent chemotherapy (recommended or administered) within 4 months (120 days) of diagnosis.

Developer Rationale: Improve the utilization of chemotherapy in women with hormone receptor negative breast cancer.

Numerator Statement: Combination chemotherapy is administered within 4 months (120 days) of the date of diagnosis or it is recommended and not received.

Denominator Statement: Women under the age of 70 with AJCC T1cN0M0, or Stage IB-III hormone receptor negative breast cancer:

- Women
- Age 18-69 at time of diagnosis
- Known or assumed first or only cancer diagnosis
- Primary tumors of the breast
- Epithelial invasive malignancy only stageable by AJCC 7th edition
- AJCC T1cN0M0, or Stage IB to III
- Primary tumor is estrogen receptor negative and progesterone receptor negative
- All or part of first course of treatment performed at the reporting facility
- Known to be alive within 4 months (120 days) of diagnosis

Denominator Exclusions: Exclude, if any of the following characteristics are identified:
Men;
Age <18 and >=70;
not a first or only cancer diagnosis;
non-epithelial and non-invasive tumors;
phyllodes tumor histology;
rare histology not supported by clinical trials: 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor, 8980 – Carcinosarcoma, 8981 - Carcinosarcoma, embryonal
Tumor size <=1cm and AJCC pN=0;
ERA positive;
PRA positive;
Evidence of in situ or metastatic disease;
Not treated surgically;
Died within 4 months (120 days) of diagnosis;
Participation in a clinical trial which directly impacts the delivery of the standard of care

Measure Type: Process

Data Source: Electronic Clinical Data : Registry, Paper Medical Records

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Mar 01, 2007 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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:

- National Comprehensive Cancer Network (NCCN) Practice Guidelines:
  - Systemic Adjuvant Treatment – Hormone Receptor- Negative- HER2- Positive Disease (Page BINV-7): pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant chemotherapy (category 1) with trastuzumab. **Level of evidence: Category 1**
  - Systemic Adjuvant Treatment – Hormone Receptor- Negative- HER2- Negative Disease (Page BINV-9): pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant chemotherapy: **Level of evidence: Category 1**
    - Category 1 is defined as: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate
- Additional evidence included a systematic review of the body of evidence including multiple randomized clinical trials demonstrating approximate 33% reduction in risk of distant cancer recurrence and death.
- The 2011 Committee expressed no concerns regarding the evidence underlying this measure

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 updated the links for the guidelines and included the NCCN Categories of Evidence and Consensus – no changes were made to the evidence.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → evidence graded as high-level evidence (Box 6) → Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on 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.

- The developer provided the following national trend data from the National Cancer Data Base (NCDB):

	2008	2013	2008-2013
# of hospitals	16,263	14,331	--
Mean Performance Rate	85.1%	89.4%	--
IQR	88-100%	92-100%	
Range	--	--	0-100%

Disparities:

Race/ethnicity

2013	Non-Hispanic white	Non-Hispanic black	Hispanic	Asian/Hawaiian/Pacific Island
# of patients	9,271	2,981	991	410
performance rates	91.5%	85.2%	83.6%	90.2%

Age

2013	Age 18-49	Age 50-59	Age 60-69
# of patients	5,119	5,043	4,169
performance rates	91.2%	89.3%	87.2%

Insurance Status

2013	Private Insurance	Medicare	Medicaid/No insurance	Other government insurance
# of patients	--	2,246	1,875	--
performance rates	91.0%	86.0%	85.0%	88.5%

- The developer provided additional disparities data on income, no high school degree, facility type, census region.
- In 2012, the Committee noted a performance gap for this measure but stated that disparities were not well documented.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Process measure. The process of administrating chemotherapy in the adjuvant setting in a timely manner is supported by the evidence and is directly related to the evidence. The evidence is long-standing, supported by high level of evidence, consistent evidence, and graded as high level by both NCCN and Early Breast Cancer Trialists Group. There is extensive documentation of the benefit of multiagent chemotherapy in women with hormone receptor negative breast cancer. Chemotherapy reduces the risk of distant disease recurrence and death by about one-third. No new evidence was provided. However, this does not interfere with the ability to evaluate the evidence since the evidence is long-standing and high quality. I rate this moderate according to the algorithm.

****The category I evidence including a systematic review for this process measure directly relates to the process measured. It relates to the desired outcome of 33% reduction in the risk of recurrence and death.**

**This is a process measure that is based on Category 1 (high evidence with uniform NCCN consensus that the intervention is appropriate. SR of RCT indicated 33% reduction in risk of distant cancer recurrence and death. No changes in the evidence since last evaluated (verified in NCCN 2016 version - BINV-7).

Do not understand why HER2+ is not clearly stated in the numerator since that is part of the NCCN guideline for this population.**

1b. Performance Gap

Comments:

Performance data was measured with updated data presented for this submission. Although there has been a steady improvement in performance, there is still a large range in performance since implementing the measure in 2008 with a persistent range from 0-100%. At the last submission, disparity gap were noted in all areas but no new data was submitted to demonstrate if this persists. An opportunity for improvement remains. I rate this moderate without complete data.

A moderate opportunity for improvement exists with EPR just below 90%. Disparities documented in race/ethnicity, insurance status, income and SES.

**Seek to improve the utilization of chemotherapy in women with ER/PR- breast cancer

Difference in 2008, 2013 in #hospitals (14,331 in 2013 and 16, 263 in 2008). MPR increased from 85.1% to 89.4%. Disparities are apparent - non-hispanic white 91.5% compared to non-hispanic black at 85.2% and hispanic at 83.6%. Uninsured/Medicaid patients have the lowest performance rating.**

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): registry and paper medical records. This is not an eMeasure.

Specifications:

- This is a facility-level measure.
- The numerator is defined as combination chemotherapy administered within 4 months (120 days) of the date of diagnosis or it is recommended and not received. Reasons combination therapy (chemotherapy or immunotherapy/BRM for patients with ER/PR negative; HER 2 Positive disease) was recommended but not administered include:
    - Contraindicated due to patient risk factors

    - Patient died prior to planned or recommended therapy
    - Recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record.
    - Recommended by the patient´s physician, but treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record.
- The denominator is defined as women under the age of 70 with AJCC T1cN0M0, or Stage IB-III hormone receptor negative breast cancer:
    - Women
    - Age 18-69 at time of diagnosis
    - Known or assumed first or only cancer diagnosis
    - Primary tumors of the breast
    - Epithelial invasive malignancy only stageable by AJCC 7th edition
    - AJCC T1cN0M0, or Stage IB to III
    - Primary tumor is estrogen receptor negative and progesterone receptor negative
    - All or part of first course of treatment performed at the reporting facility
    - Known to be alive within 4 months (120 days) of diagnosis
- Denominator exclusions include:
    - Men;
    - Age <18 and >=70
    - not a first or only cancer diagnosis
    - non-epithelial and non-invasive tumors
    - phyllodes tumor histology
    - rare histology not supported by clinical trials
    - Tumor size <=1cm and AJCC pN=0
    - ERA positive
    - PRA positive
    - Evidence of in situ or metastatic disease
    - Not treated surgically
    - Died within 4 months (120 days) of diagnosis
    - Participation in a clinical trial which directly impacts the delivery of the standard of care - this is an update since last endorsement date
- A calculation algorithm is provided.
- All cases which meet the measure criteria are included in the denominator. If a required data element is missing the case is flagged for additional review.
- Diagnosis codes are based on the Facility Oncology Registry Data Standards (FORDS), which were revised in 2016. Therefore, no ICD-9 or ICD-10 codes are provided for this measure.
- The database is a hospital cancer registry reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base.
- In 2012, the Committee noted that the measure did not specify that the patient received the gold standard for combination chemotherapy; as such, patients could be getting less mainstream combination chemotherapy and that would still count toward the numerator. The Committee also question how neoadjuvant chemotherapy was captured. The developer clarified that the date of service of the chemotherapy and the clinical and pathological staging were all captured. The Committee expressed a desire to see a more nuanced iteration of the measure in the future to capture whether the chemotherapy administered was appropriate.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 1,400 cancer programs and approximately 14,000 cases from from all CoC-accredited cancer programs. The mean performance rates across all CoC-accredited cancer programs in 2007 was 86.3 and 84.9 in 2008. Cancer programs in the 75th percentile had performance rates of 100 in each respective year; 6.4% of programs had statistically low outlier performance rates (<44%), SD=22.8%.
- In 2012, the Committee questioned what an acceptable performance rate for the measure is. The developer stated that the target rate is 90 percent, knowing that there should be some flexibility. It was also noted that this measure captures consideration of or administration of combination chemotherapy, making it somewhat easier to achieve the numerator.

Describe any updates to testing:

- The developer provided updated data from 2013:
  - Minimum hospital-level performance rate: 0%
  - Maximum hospital-level performance rate: 100%
  - 428 programs reported Estimated Performance Rates (EPR) in the lowest quartile ≤89.5% with the lowest decile reporting EPRs of ≤75.0%

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Overall performance rates do not meet criterion.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability : Precise specifications (Box 1) →empirical testing as specified (Box 2) → empirical validity testing at patient level (Box 3)→use rating from validity testing of patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

Questions for the Committee:

- *See questions under Validity*

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Randomly selected charts were reviewed by site surveyors to determine completeness and validity of data reported to registry. The measure denominator and numerator were viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Describe any updates to testing: The developer provided additional details on data element validity testing - see below

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**

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

Validity testing method:

- The developer provided the following information about the dataset: Survey sites and data collection occurred in 2009 and 2010. In 2009, 391 sites were reviewed and 5,712 charts - 5,390 of these charts were breast cases from 2006; representing 15.8% of measure eligible cases. In 2010, 423 sites were reviewed and 6,752 charts – 6,370 of these charts were breast cases from 2007; representing 15.7% of measure eligible cases.
- Data elements reviewed:
  - confirmation of timing of adjuvant therapy
  - documentation of treatment recommended but not received
  - assessment of missing and incomplete tumor characteristics

Validity testing results:

- The developer provided the following testing results:
  - "Assessment of timing for chemotherapy for cases in which treatment was significantly early this measure had the highest concordance with 81.1 in 2006 diagnoses and 75.7% for 2007 cases. There was 88.1% and 89.5% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure. A total of 298 cases with missing hormone receptor status were reviewed, this information was found in nearly 90% of these cases."
- The developer provided percentage agreement results for two of the data elements included in the numerator (timing of chemotherapy and chemotherapy recommended but not received). NQF guidance states that testing should be done for all critical data elements.
- Site surveyors determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.
- Developers provided only percentage agreement statistics. The results reported for the data element 'timing for chemotherapy' for 2006 (81.1 %) and 2007 (75.7%) are concerning; no additional results were provided (e.g., kappa scores, which indicate agreement over and above chance; sensitivity or specificity statistics).

Questions for the Committee:

- *Does the measure adequately identify and include colon cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

Exclude, if any of the following characteristics are identified:

- Men;
- Age <18 and >=70;
- not a first or only cancer diagnosis;
- non-epithelial and non-invasive tumors;
- phyllodes tumor histology;
- rare histology not supported by clinical trials: 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor, 8980 – Carcinosarcoma, 8981 - Carcinosarcoma, embryonal
- Tumor size <=1cm and AJCC pN=0;
- ERA positive;
- PRA positive;
- Evidence of in situ or metastatic disease;
- Not treated surgically;
- Died within 4 months (120 days) of diagnosis;
- Participation in a clinical trial which directly impacts the delivery of the standard of care

- The measure exclusions as described are the opposite of the measure inclusion criteria. The cases excluded are those in which the clinical evidence does not support inclusion in the quality measure.
- In 2012-2013, 1 case (<0.01%) was excluded due to patient participating in clinical trial that directly impacts delivery of the standard of care; this exclusion does not affect estimated performance rates for this measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Performance data is presented above under opportunity for improvement. Complete details of the data are presented in 1b.

Question for the Committee:

- *Given the data provided in 1b., does the measure identify meaningful differences about quality across facilities?*

2b6. Comparability of data sources/methods:

- Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b7. Missing Data

- The developer describes in S.22 that all cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review. The developer does not provide information on the frequency of missing data or potential impact on results.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) → validity testing conducted with patient-level data elements (Box 10)→ Only assessed percent agreement for two data elements in numerator (Box 11)→Insufficient

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

Rationale: Validity testing results for all critical data elements not presented and percent agreement results alone do not meet data-element validity criterion.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The specifications are clearly defined and appropriate. No concerns about implementing the measure.
Data elements are clearly defined, percent agreement provided for two data elements and so the measure does not meet data-element criterion. No significant concern about the likelihood of the measure not being implemented consistently.
Some changes in numerator elements were made: Recommended replaced "considered". Rare histologies were deleted (not supported by evidence), and included in the exclusion criteria women involved in clinical trials that impact SOC.
The specifications are consistent with the evidence. They are appropriate to use for this measure.
Specifications are consistent with the evidence.

2a2. Reliability Testing

Comments:

The reliability algorithm suggests that the testing is insufficiently reliable and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff. I rate this insufficient.
Sufficient scope to generalize for widespread implementation. Overall performance does not meet criterion for reliability testing.
A calculation algorithm was provided. Updates - in 2013-428 programs reported EPR in the lowest quartile less than or equal to 89.5 with the lowest less than or equal to 75. The overall performance rates did not meet criterion - insufficient.

2b2. Validity Testing

Comments:

The validity algorithm suggests that the testing is insufficiently valid and does not meet the standards provided by the NQF. I am not certain how the committee should interpret this. I recommend further discussion and guidance from the NQF staff. At this point, the measure does not pass the reliability and validity standards as set forth by NQF. I rate this insufficient.
Percent agreement provided for two data elements and so the measure does not meet data-element criterion. Results for "timing for chemotherapy" noted as concerning.
Data element validity testing counts for data element reliability. The specifications are consistent with the evidence (not sure about the lack of clearly stating HER2+ in the numerator). Two data elements had percentage agreements - not all elements were tested. No other statistics included (e.g., kappa scores). Viewed as insufficient based on NQF criteria

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias

Comments:

**The exclusions are appropriate to define the correct population. There is no risk adjustment. It appears that the analyses can demonstrate meaningful differences since there is a standard methodology across all sites and a

credentialing process for the sites. The analyses should allow for comparisons. However, in all cases, there is a plan for missing data but further information regarding the success of acquiring this data is not provided with this submission.**
Exclusions are consistent with the evidence and groups are not inappropriately excluded. No risk adjustment. Higher EPR indicate improvement in care related to desired health outcome.
1 case was excluded due to patient participating in clinical trial that impacts SOC, but did not affect estimated PR. However, the frequency of missing data or potential impact on results is not reported.

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.

- Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry); some data elements are in defined fields in electronic sources.
- Data collection burden due to manual chart abstraction from paper medical records.

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

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
All of the data is routinely used during patient care and should be available in the medical record. The majority of the data requires abstraction which is fraught with potential area and burdensome to obtain. I have no concerns about the data collection strategy. I rate this moderate.
Data elements would typically be generated in EHR's although may vary in evolving environment. Use of CTRs presents some level of burden if manual data collection is necessary.
Abstracted from electronic sources and paper medical records by someone other than person obtaining original information.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:
- Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.
- The National Cancer Data Base (NCDB) provides a venue for accredited programs to benchmark their compliance

compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP). CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer. This application is available to over 1500 CoC-accredited cancer programs. CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

- In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.
- QOPI® Certification Program (adapted):The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score.

Improvement results :

- The developer provided the following improvement results:
  - 2008: 85.1 (84.6 – 85.7); n=16,151
  - 2009: 87.8 (87.3 – 88.3); n=15,835
  - 2010: 91.3 (90.9 – 91.8); n=15,731
  - 2011: 91.2 (90.8 – 91.7); n=15,761
  - 2012: 90.5 (90.0 – 91.0); n=15,043
  - 2013: 89.5 (89.0 – 90.0); n=14,264

Unexpected findings (positive or negative) during implementation:

- This measure, as specified, is susceptible to under-reporting of the adjuvant hormone therapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. However, CoC accredited programs have demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in higher and clinically more accurate reflections of the care provided or coordinated through their centers. It does take additional time to collect and report this adjuvant therapy information. Additionally, the CoC´s Program Standards require review of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

Potential harms:

- Developer did not identify any unintended consequences related to this measure.

Feedback :

- Developer did not identify any specific feedback loops related to this measure.

***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

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
This measure is being publically reported and is used in multiple settings. Since the timely administration of adjuvant therapy has a large impact on survival, this measure can be helpful in improving health care. No unintended consequences. The benefits of this measure outweigh the risks. I rate usability high.
Pennsylvania Health Care Quality Alliance. Quality Oncology Practice Initiative.
This measured is used by several entities and publically reported (e.g., CoC, QOPI, PHCQA).

Criterion 5: Related and Competing Measures
Related or competing measures N/A **Harmonization**

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0559 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)***:**

Process

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Directly applicable - randomized trials examining the measure

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles*)**:** Multiple randomized clinical trials

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* High quality evidence

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Strong level of consistency

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

Approximate 33% reduction in risk of distant cancer recurrence and death

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** National Comprehensive Cancer Network (NCCN): Early Breast Cancer Trialists Collaborative Group

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Level I, IIA , IIB, III

1c.13 **Grade Assigned to the Body of Evidence:** Level 1

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

See 1.b4

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:

Systemic Adjuvant Treatment – Hormone Receptor- Negative- HER2- Positive Disease (Page BINV-7)

pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant chemotherapy (category 1) with trastuzumab (category 1)

Systemic Adjuvant Treatment – Hormone Receptor- Negative- HER2- Negative Disease (Page BINV-9)

pT1, pT2, or pT3; and pN0 orpN1mi –and pN0 orpN1mi ->Tumor >1cm -> Adjuvant chemotherapy (category 1)

1c.17 Clinical Practice Guideline Citation: NCCN Clinical Practice Guidelines v1.2016

1c.18 National Guideline Clearinghouse or other URL: http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: National Comprehensive Cancer Network (NCCN)

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: Level I, IIA , IIB, III

1c.23 Grade Assigned to the Recommendation: Level 1

1c.24 Rationale for Using this Guideline Over Others: All guidelines recommend chemotherapy with hormone receptor negative breast cancer

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0559_Evidence_MSF5.0_Data.doc,MAC_0559_Evidence_2016-635953599729162351.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Improve the utilization of chemotherapy in women with hormone receptor negative breast cancer.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
The nationally recognized National Cancer Data Base (NCDB), jointly sponsored by the American College of Surgeons and the American Cancer Society, is a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent approximately 70 percent of newly diagnosed cancer cases nationwide and 30 million historical records. Data from the NCDB was analyzed.

The NCDB collects data from CoC accredited cancer programs on an annual basis; the data we collect is in accordance with standard registry procedures. In January of 2015, 2013 diagnoses were collected. This information was released to accredited cancer programs in the late summer and is included in this applications.

The mean performance rate for this measure has increased from 85.1% (95% CI: 84.5-85.6) IQR=88-100% n= 16,263 in 2008 to 89.4% (88.9-89.9) IQR=92-100% n=14,331 in 2013 representing a steady improvement in quality. The minimum hospital-level performance rate is 0% with a 100% maximum in all years assessed 2008-2013.

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.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
The data source is described in 1b.1. Disparities were assessed by race/ethnicity, age, insurance status, facility type, and education and income at the zip code level.

Race/ethnicity
Race/ethnicity was defined as non-Hispanic white, non-Hispanic black, Hispanic, Asian/Pacific Island or other race/ethnicity. Between 2008 and 2012, performance rates increased in all ethnic groups. Non-hispanic whites had the highest performance rates in 2013 at 91.5% (95% CI: 91.0-92.1) n=9271 in 2013, followed by Asian Pacific Islanders 90.2% (87.3-93.1) n=410, Blacks 85.2% (83.9-86.4) n=2981, and Hispanics 83.6% (81.3-85.9) n=991. Between 2008 and 2012, performance rates increased in all ethnic groups.

Age
Age groups were defined as, 18-49, 50-59, 60-69. Since 2008, each age group saw a relatively equal gain in performance with the measure. Patients under the age of 50 at diagnosis had marginally higher performance rates in 2013 at 91.2% (90.4-91.9) n=5119, compared to patients 50 to 59 89.3% (88.5-90.2) n=5043, and 87.2% (86.2-88.2) n=4169.

Insurance Status
Insurance status is defined as insurance at the time of diagnosis. Insurance was stratified into private, Medicare, Medicaid/ No insurance. Since 2008, patients with each insurance type saw a gain in performance. Uninsured and Medicaid patients had the lowest performance rates in 2013 at 85.0% (83.3-86.6) n=1875, Medicare at 86.0% (84.6-87.4) n=2246, Other Government 88.5% (84.7-92.4), with private insurance having the highest performance rates at 91.0% (90.4-91.6).

Median Income Quintile
Income quintiles at the zip code level were assessed based on the 2012 American Community Survey. Patients that resided in communities with a median income of <$36,000 annually at diagnosis experienced lower performance in 2013 than patients from communities with a median income above $36,000. In 2008, the mean performance rate for <$36K was 83.1% (81.5-84.6) n=2246 and increased only 2.2% by 2013 to 85.3% (83.7-86.8) n=2009. In contrast, patients that resided in communities with median incomes above $36K experienced a 5% gain in performance between 2008 and 2013 from 85.0% to 90.0%.

SES – Proportion of population with no high school degree in patient zip code

The proportion of the population with no high school degree at the zip code level were assessed based on the 2012 American Community Survey. Patients that resided in communities at time of diagnosis with the lowest proportion of no high school degree (<7%) had higher rates of performance in 2013 91.5% (90.6-92.4) n=3534) than patients from communities with the highest proportion of patients with no high school degree (>21%) 85.3% (83.9-86.7) n=2535. Likewise, the performance increase from 2008 was smaller for patients from communities with the lowest proportion of no high school degree (<7%) 3.4% gain compared to zip codes with great proportions of residents without a high school degree.

Facility Type

Facility type was assessed by CoC-accreditation status; facility types include Comprehensive Community Cancer Programs, Integrated Network Cancer Programs, Community Cancer Programs and by Teaching/Research programs. Patients that were treated at teaching/research hospitals experienced similar performance rates 89.3% (88.4-90.1) n=5087 to those treated at comprehensive community centers 89.8% (89.0-90.5) n=6382 in 2013. Patients treated at smaller community hospitals had only slightly lower performance rates at 87.1% (85.4-88.7) n=1529. Since 2008, patients at academic hospitals experienced the largest gain in performance (+8.7%), whereas the performance rate in 2008 for patients at community centers was 87.3% (86.5-88.0), representing only a 2.5% gain. The same trend is true for community hospitals, with performance at 85.4% (83.8-87.0) in 2008 representing a 1.7% gain.

Census region

Performance rates increased in all census regions between 2008 to 2013. Patients that resided in the Northeast Census Region at time of diagnosis experienced the largest gain in performance between 2008 and 2013 compared to all other Census regions. In 2008, the average performance rate for the Northeast was 77.7% (76.2-79.2) n=3039 in contrast to 2013 where it had risen 11% to 88.7% (87.6-89.9) n=2865. In 2013, patients residing in the Midwest at time of diagnosis had the highest performance rate at 91.6% (90.7-92.5%) n=3539. The West had a 2013 performance rate of 89.9% (87.4-92.3%) n=595, the South had a 2013 performance rate of 89.0% (88.1-89.8) n=5775, the Pacific had a 2013 performance rate of 87.4% (85.7-89.0) n=1508.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

There is extensive documentation of the benefit of mutiagent chemotherapy in women with hormone receptor negative breast cancer. Chemotherapy reduces the risk of distant disease recurrence and death by about one-third. The restriction to women under age 70 is because this measure is for the purpose of provider accountability. There are limited data in women over age 70 to guide recommendations and a higher fraction of these women have reasons to omit chemotherapy including co-morbidity.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Early Breast Cancer Trialists Collaborative Group (EBCTCG) et al. Comparisons between different polychemotherapy regimens for early breast cancer: metaanalysis of long-term outcome among 100,000 women in 123 randomised trials. Lancet 2012;379(9814):4320444. 2. Early Breast Cancer Trialists Collaborative Group (EBCTCG) et al. Adjuvant chemotherapy in oestrogen-receptor-poor breast cancer: patient level meta-analysis of randomised trials. Lancet 2008;371(9606):29-40. 3. Early Breast Cancer Trialists Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: and overview of the randomised trials. Lancet 2005;365(9472):1687-1717. 4. Hassett MJ, Hughes ME, Niland JC, et al. Selecting high priority quality measures for breast cancer quality improvement. Med Care 2008;46:762-770.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Breast

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Disparities

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.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Since the last endorsement maintenance minor changes to this measure have been instituted.

The word considered has been replaced in the numerator statement with recommended to be more consistent with the registry codes used to assess this measure.

Rare histologies which are not supported by clinical evidence were removed from inclusion:
8200 - adenoid cystic carcinoma,
8940 - Mixed tumor, malignant, NOS
8950 - Mullerian mixed tumor
8980 - Carcinosarcoma
8981 - Carcinosarcoma, embryonal
9020- phyllodes tumor

Based on changes in SEER coding of chemotherapy and immunotherapy patients with HER2 positive disease diagnosed after 2013 are compliant with the standard if they receive chemotherapy plus Her2 targeted therapy (immunotherapy), or if these treatments are recommended and not received.

An exclusion to remove patients in which participation in a clinical trial which directly impacts the delivery of the standard of care.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Combination chemotherapy is administered within 4 months (120 days) of the date of diagnosis or it is recommended and not received.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
4 months (120 days)

S.6. **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, 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.
Chemotherapy [NAACCR Item#1390]=3, and Date Chemotherapy Started (NAACCR Item#1220] <=120 days following Date of Diagnosis [NAACCR Item# 340] OR
Chemotherapy recommeneded and not received [NAACCR Item#1390]=82-87 (82:not recommended/administered because it was contraindicated due to patient risk factors, 85:not administered because the patient died prior to planned or recommended therapy,86:It was recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record. 87: it was recommended by the patient´s physician, but this treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record) OR;
OR
For patients ER/PR negative; Her2 Positive disease
Chemotherapy [NAACCR Item#1390]=2,3 and Date Chemotherapy Started (NAACCR Item#1220] <=120 days following Date of Diagnosis [NAACCR Item# 340]
AND
Immunotherapy/BRM recommended and not received [NAACCR Item#1410]=82-87 (82:not recommended/administered because it was contraindicated due to patient risk factors, 85:not administered because the patient died prior to planned or recommended therapy,86: recommended by the patient´s physician, but was not administered as part of first-course therapy. No reason was stated in the patient record. 87: recommended by the patient´s physician, but this treatment was refused by the patient, the patient´s family member, or the patient´s guardian. The refusal was noted in the patient record)
OR; Immunotherapy/BRM [NAACCR Item#1410]=1 and Date Immunotherapy Started (NAACCR Item#1240] <=120 days following Date of Diagnosis [NAACCR Item# 340]

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women under the age of 70 with AJCC T1cN0M0, or Stage IB-III hormone receptor negative breast cancer:

- Women
- Age 18-69 at time of diagnosis
- Known or assumed first or only cancer diagnosis
- Primary tumors of the breast
- Epithelial invasive malignancy only stageable by AJCC 7th edition
- AJCC T1cN0M0, or Stage IB to III
- Primary tumor is estrogen receptor negative and progesterone receptor negative
- All or part of first course of treatment performed at the reporting facility
- Known to be alive within 4 months (120 days) of diagnosis

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Sex [NAACCR Item#220]=2; Age at Diagnosis [NAACCR Item#230] =18-69; Sequence number [NAACCR Item # 560] = 00-01;
Tumor Size [NAACCR Item#2800]= 011-898, 992-995 and AJCC pN [NAACCR Item#890]=0,0i-, 0I=, 0M-, 0M+; OR AJCC pN [NAACCR

Item#890]=1,1a, 1b, 1c, 2,2a, 2b, or 3, 3a, 3b,3c; AND CS SSF1 (ERA) [NAACCR Item#2880]=020, 30; AND CS SSF2 (PRA) [NAACCR Item#2890]=020 or 030; AND Surgical Procedure of the Primary Site [NAACCR Item#1290] = 20–90

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Exclude, if any of the following characteristics are identified:
Men;
Age <18 and >=70;
not a first or only cancer diagnosis;
non-epithelial and non-invasive tumors;
phyllodes tumor histology;
rare histology not supported by clinical trials: 8940 - Mixed tumor, malignant, NOS, 8950 - Mullerian mixed tumor, 8980 – Carcinosarcoma, 8981 - Carcinosarcoma, embryonal
Tumor size <=1cm and AJCC pN=0;
ERA positive;
PRA positive;
Evidence of in situ or metastatic disease;
Not treated surgically;
Died within 4 months (120 days) of diagnosis;
Participation in a clinical trial which directly impacts the delivery of the standard of care

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No stratification applied

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
NA

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
NA

S.16. Type of score:
Rate/proportion
If other:

S.17. 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.18. Calculation Algorithm/Measure Logic *(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; aggregating data; risk*

adjustment; etc.)
This measure score is calculated by dividing the numerator cases by denominator eligible cases.

Denominator eligible cases are assessed in a step-wise fashion:
- Include breast cancer case
- Exclude patients enrolled in a clinical trial that directly impacts the delivery of the standard of care.
- Include female patients only
- inlcude patients 18-69
- Include epithelial tumors which can be staged according to the AJCC 7th Ed (8000-8199, 8201-5876, 8941-8949)
- Include invasive tumors only
- Exclude patients with pathologic evidence of in situ or metastatic disease
- Exclude patients with clinical evidence of in situ or metastatic disease
- Include cases where all or part of the first course of treatment was performed at the reporting facility
- Include only surgically treated cases
- Includes patients reported living withing 120 days from diagnosis
- Include AJCC T1cN0M0 or AJCC Stage IB -III tumor
- Hormone receptor negative cases

Numerator cases are then assessed from denominator eligible cases:
- Cases with HER2 negative disease: Combination chemotherapy administered within 120 following diagnosis or Chemotherapy recommended but not administered
- Cases with HER2 positive disease: Chemotherapy and Her2 targeted therapy (immunotherapy) both administered within 120 days following diagnosis or chemotherapy administered within 120 days and Her 2 targeted therapy (immunotherapy) recommended

The measure score is calculated with the numerator divided by the denominator.

See: https://www.facs.org/~/media/files/quality%20programs/cancer/quality%20breast.ashx

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above 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. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
NA

S.21. Survey/Patient-reported data *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
NA

S.22. Missing data (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
All cases which meet the measure criteria are included in the denominator. If a required data element is missing; the case is flagged for additional review.

S.23. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry, Paper Medical Records

S.24. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospital cancer registry data, reported to the American College of Surgeons, Commission on Cancer, National Cancer Data Base. Data is collected in accordance with the North American Association of Central Cancer Registries (NAACCR) coding

http://www.naaccr.org/Applications/ContentReader/Default.aspx

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0559_MeasureTesting_MSF5.0_Data.doc,0559_MeasureTesting_MAC_04012016.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0559 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. **Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

This measure has been implemented by the ACoS CoC since 2007 across all CoC-accredited cancer programs, and reports on approximately 14,000 cases per year to almost 1,400 cancer programs.

2a2.2 **Analytic Method** *(Describe method of reliability testing & rationale)***:**

Cancer registry case records reported to the NCDB are reviewed annually, annualized hospital performance rates are provided back to CoC accredited cancer programs via the CoC´s Cancer Program Practice Profile Report (CP3R) using the denominator and numerator criteria documented in response to items 2a1.3 and 2a1.7, respectively, in the Specifications section.

(http://www.facs.org/cancer/ncdb/cp3r.html)

2a2.3 **Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

The mean performance rates across all CoC-accredited cancer programs was 86.3 in 2007 and 84.9 in 2008. The two years available at the time of this writing. Cancer programs in the 75th percentile had performance rates of 100 in each respective year. Even with high aggregate performance rates demonstrated by programs room for improvement across the system of CoC-accredited programs remains, with 6.4% of programs with statistically low outlier performance rates (<44%), SD=22.8%.

In 2013, the minimum hospital-level performance rate is 0% with a 100% maximum. 428 programs reported EPRs in the lowest quartile ≤89.5% with the lowest decile reporting EPRs of ≤75.0%.

2b. VALIDITY. Validity, Testing, <u>including all Threats to Validity</u>: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

See 2a2.1. This measure has been implemented across all CoC-accredited cancer programs and subject to local review by standing committees of these hospitals and site surveyors at the time of accreditation site visits.

During Commission on Cancer Survey Site visits in 2009 and 2010, surveyors validated not more than 25 charts.

During 2009 – 391 accredited sites were reviewed, including 5,712 charts. This included an average of 14 charts per survey (IQR 6-22). 5,390 of these charts were breast cases; representing 15.8% of measure eligible cases.

During 2010- 423 accredited sites were reviewed, including 6,752 charts. This was based on an average of 14 charts per survey (IQR 6 – 22). 6370 of these charts were breast cases; representing 15.7% of measure eligible cases.

2b2.2 **Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

Performance rates are reviewed and discussed, randomly selected charts are reviewed by the site surveyor to ascertain the completeness and validity of the data recorded in the local cancer registry and reported to the NCDB and included in the CP3R reporting application.

Major areas of review completed by site surveyors included but were not limited to, confirmation of timing of adjuvant therapy, documentation of treatment recommended but not received, assessment of missing and incomplete tumor characteristics.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure has a high degree of user acceptability, the measure denominator and numerator are viewed by the clinical constituency within these cancer programs as valid and an appropriate reflection of the standard of care described in NCCN clinical guidelines.

Assessment of timing for chemotherapy for cases in which treatment was significantly early this measure had the highest concordance with 81.1 in 2006 diagnoses and 75.7% for 2007 cases. There was 88.1% and 89.5% agreement in 2006 and 2007 diagnoses respectively for chemotherapy which was recommended but not administered for this measure. A total of 298 cases with missing hormone receptor status were reviewed, this information was found in nearly 90% of these cases.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

The NCDB collects all diagnosed cases within cancer programs. The measure exclusions as described in the specifications are the inverse of the measure inclusion criteria. Measure exclusions are based on parameters in which the clinical evidence does not support inclusion in the quality measure. These are established to ensure patients included in the measure assessment meet the evidence based criteria. In 2012 14,331 breast cases were included in this measure.

The exception to this is the measure exclusion of, "Patient participation in a clinical trial which directly impacts the standard of care."

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

An assessment of cases using the measure exclusion for "Participation in a clinical trial which directly impacts the standard of care' was reviewed. For all cases applicable to this measure, in 2012 -2013, 1 case was excluded from the measure denominator based on the exclusion of patient participation in a clinical trial which directly impacts the standard of care.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

Measure exclusions were used for n=1 (<0.01%) and does not affect estimated performance rates for this measure.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Differences in performance were described in the application

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Not applicable; all data are reported from CoC-accredited programs, collected in a standardized fashion and reported via the standard NAACCR record transmission layout.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:**

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**

This measure was not specified to report stratified performance rates, however the CoC´s recently released (2011) "real clinical time" Rapid Quality Reporting System (RQRS) (http://www.facs.org/cancer/ncdb/rqrs.html) reports back measure-specific performance rates by a number of strata, eg. patient age, sex, ethnicity, insurance status, and area-based SES. RQRS hosts a prospective treatment alert system, and so performance rates are both high and consistent with clinical expectation, however room for potential improvement remains. In a comparative analysis of 16 NCI/NCCCP pilot sites using RQRS with a comparative group of 25 other CoC-accredited cancer programs also using RQRS revealed that across all 41 hospitals 88.1-88.5% of patients (white or African-American) were concordant with receipt of multi-agent chemotherapy, and that 88.9-90.4% of patients, based upon income SES were also concordant, without regard to income status or reporting hospital. Analysis from cases diagnosed 2008-2010.

Additional disparities data was presented in section 1.b. of this application.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

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)

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)*

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.**

The ACoS/CoC implementation of this measure is framed around the feasibility of data collection and reporting considerations. Cancer registries in the United States depend on a multitude of information sources in order to completely abstract case records and be in compliance with State, Federal and private sector accreditation requirements. Commission on Cancer Standards require case abstracting to be performed by a Certified Tumor Registrars (CTRs). CTRs must pass an exam and maintain continuing education. In the past decade, great strides have been made within the cancer registration community in terms of electronic capture of registry data from electronic pathology systems and electronic health records. However, until EHR systems are universally implemented in the US and fully integrated within hospital-level cancer registry systems, registry data will depend upon some level of human review and intervention to ensure data are complete and accurately recorded. Robust data quality edits are applied to the data at all levels of cancer data abstraction and processing. These edits standardize coded information and ensure its accuracy.

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.**

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The infrastructure to monitor compliance with this measure has been in place since 2005 to assess and feed-back to approximately 1,500 Commission on Cancer (CoC) accredited centers performance rates for this measure. CoC accredited cancer programs account for 70-80% of patients affected by this measure. This measure is currently reported to CoC accredited programs through the National Cancer Data Base (NCDB) using the Cancer Program Practice Profile Report (CP3R) web-based audit and feed-back reporting tool. The CP3R is generally described at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. In addition, this measure is also reported to over 1030 cancer programs participating in its "real clinical time" feedback reporting tool through its Rapid Quality Response System (RQRS). An overview of the RQRS is available at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Both of these reporting tools have been utilized in the cancer registry community and do not produce an undue burden on the data collection network. Utilization of these tools increases the completeness of adjuvant therapy information captured by the cancer registry.

The data for this measure are key elements already collected in all hospital registries. This measure has been reviewed using cancer registry data. The CoC data demonstrates variation in the measure. Registries have demonstrated the ability to identify gaps in data collection and to correctly identify therapy in the majority of cases. The measure is readily implemented.

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)*.

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Pennsylvania Health Care Quality Alliance http://www.phcqa.org/ https://www.medicare.gov/hospitalcompare/cancer-measures.html PPS-Exempt Cancer Hospital Quality Reporting program

	Professional Certification or Recognition Program http://www.instituteforquality.org/qcp/qopi-certification-measures QOPI® Certification Program Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/manual-qopi-measures Commission on Cancer, National Cancer Data Base https://www.facs.org/quality%20programs/cancer/ncdb

4a.1. **For each CURRENT use, checked above, provide:**

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

a) Public Reporting
Pennsylvania Health Care Quality Alliance
Purpose: The Pennsylvania Health Care Quality Alliance (PHCQA) is a voluntary group of health care organizations collaboratively working together to improve the quality of health care for the people of Pennsylvania. The PHCQA allows for voluntary reporting of compliance with CoC Measures by accredited programs in the state, currently 60 of 71 eligible programs participate.

f) Quality Improvement with Benchmarking

Commission on Cancer, National Cancer Data Base
Purpose: The National Cancer Data Base (NCDB) provides a venue for accredited programs to benchmark their compliance compared to other CoC-accredited cancer programs through the use of the Cancer Program Practice Profile Reports (CP3R), the Rapid Quality Reporting System (RQRS) and the Cancer Quality Improvement Program (CQIP).

CP3R offers local providers comparative information to assess adherence to and consideration of standard of care therapies for major cancer and is described https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cp3r. This application is available to over 1500 CoC-accredited cancer programs

CQIP reports annual quality and outcomes data to more than 1,500 cancer programs accredited by the American College of Surgeons Commission on Cancer (CoC) and provides the availability for programs to benchmark their performance on quality measures to other CoC-accredited programs. https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/cqip

RQRS is a reporting and quality improvement tool which provides real clinical time assessment of hospital level adherence to National Quality Forum (NQF)-endorsed quality of cancer care measures for breast and colorectal cancers - See more at: https://www.facs.org/quality-programs/cancer/ncdb/qualitytools/rqrs. Over 1040 CoC-accredited cancer programs across the country are currently participating in this quality tool.

Quality Oncology Practice Initiative (adapted):
In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

QOPI® Certification Program (adapted):
The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score. Please see a description of the QOPI® program above for details.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

2008: 85.1 (84.6 – 85.7); n=16,151
2009: 87.8 (87.3 – 88.3); n=15,835
2010: 91.3 (90.9 – 91.8); n=15,731
2011: 91.2 (90.8 – 91.7); n=15,761
2012: 90.5 (90.0 – 91.0); n=15,043
2013: 89.5 (89.0 – 90.0); n=14,264

Patient demographics, including ethnicity and age were considered. The trends reveal an increase in performance across all ethnicity and age groups, although the rates within vary. Additional variation exists in the rate of increase from 2008 to 2013. Non-hispanic whites had the highest performance rates in 2013 at 91.5% (95% CI: 91.0-92.1) n=9271 in 2013, followed by Asian Pacific Islanders 90.2% (87.3-93.1) n=410, Blacks 85.2% (83.9-86.4) n=2981, and Hispanics 83.6% (81.3-85.9) n=991. Between 2008 and 2013, performance rates increased in all ethnic groups. Patients under the age of 50 at diagnosis had marginally higher performance rates in 2013 at 91.2% (90.4-91.9) n=5119, compared to patients 50 to 59 89.3% (88.5-90.2) n=5043, and 87.2% (86.2-88.2) n=4169. Since 2008, each age group saw a relatively equal gain in performance with the measure.

Geographic variation exists. Patients that resided in the Northeast Census Region at time of diagnosis experienced the largest gain in performance between 2008 and 2013 compared to all other Census regions. In 2008, the average performance rate for the Northeast was 77.7% (76.2-79.2) n=3039 in contrast to 2013 where it had risen 11% to 88.7% (87.6-89.9) n=2865. In 2013, patients residing in the Midwest at time of diagnosis had the highest performance rate at 91.6% (90.7-92.5%) n=3539. The West had a 2013 performance rate of 89.9% (87.4-92.3%) n=595, the South had a 2013 performance rate of 89.0% (88.1-89.8) n=5775, the Pacific had a 2013 performance rate of 87.4% (85.7-89.0) n=1508.

4b.2. **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.**

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
This measure, as specified, is susceptible to under-reporting of the adjuvant chemotherapy component appearing in the measure numerator. Due to referral of services, access to patient clinical follow-up with radiation oncology may initially be limited or unavailable. Programs use of the CoC data quality tools has demonstrated through retrospective case and chart reviews that significant additional and accurate information regarding treatment provided to patients can be ascertained, resulting in more accurate reflections of the care provided or coordinated through their centers. Additionally, the CoC´s Program Standards require

direct review and oversight of quality measures be monitored by an attending physician (Cancer Liaison Physician) on staff at the center on a quarterly basis.

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

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 completely harmonized?

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):** Commission on Cancer, American College of Surgeons
Co.2 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-
Co.3 **Measure Developer if different from Measure Steward:** Commission on Cancer, American College of Surgeons
Co.4 **Point of Contact:** Erica, McNamera, emcnamara@facs.org, 302-202-5194-

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 developers:
Christopher Pezzi, MD, FACS (Abington Memorial Hospital, Abington PA); Lawrence Shulman, MD (Dana Farber Cancer Institute, Boston MA); Stephen Edge, MD, FACS (Roswell Park Cancer Institute, Buffalo NY); David Winchester, MD, FACS (Northshore University Health System, Evanston IL); Diana Dickson-Witmer, MD, FACS (Chistiana Health Care System, Wilmington DE); Kelly Hunt, MD, FACS (MD Anderson Cancer Center, Houston TX); Marilyn Leitch, MD, FACS (University of Texas – Southwestern, Dallas TX); Katherine Virgo, PhD (American Cancer Society)

The current Measure workgroup includes:
Charles Cheng MD, FACS (Fox Valley Surgical Associates, Appleton, WI), Daniel McKellar, MD, FACS (Wayne Healthcare, Greenville, OH), David Jason Bentrem, MD (Northwestern Memorial Hospital, Chicago, IL), Karl Bilimoria, MD, FACS (Northwestern Univ/Feinberg Sch of Med, Chicago, IL), Lawrence Shulman MD (University of Pennsylvania, Philadelphia, PA), Matthew A Facktor, MD FACS (Geisinger Medical Center, Danville, PA), Ted James (University of Vermont, Burlington, VT)

This panel meets at least once annually to review quality measures currently supported and implemented by the ACoS Commission on Caner and to investigate and consider/review development of possible new measures.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 11, 2015
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?** 11, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 1857
Measure Title: HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies
Measure Steward: American Society of Clinical Oncology
Brief Description of Measure: Proportion of female patients (aged 18 years and older) with breast cancer who are human epidermal growth factor receptor 2 (HER2)/neu negative who are not administered HER2-targeted therapies
Developer Rationale: Human epidermal growth factor receptor (HER2) gene is amplified and/or overexpressed in approximately 15% to 20% of primary breast cancers (Giordano, 2014). The ASCO/CAP joint guideline on HER2 testing recommends all patients with invasive breast cancer should be tested for HER2 status and only those who test positive for HER2 status should receive HER2 targeted therapies. Additionally data have shown that the administration of HER2 targeted therapies such as Pertuzumab offer no clinical benefit in patients with HER2 negative metastatic disease (Wolff, 2013).

The contraindicated administration of HER2 targeted therapy to patients with HER2 negative breast cancer can propagate potentially toxic, costly and adverse effects as well as decrease the patient's overall quality of life (Partridge, 2014).

Citations:

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Partridge, A.H., Smith, I.E., et. al., "Chemo- and Targeted Therapy for Women with Human Epidermal Growth Factor Receptor 2- Negative (or Unknown) Advanced Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Onc Pr 11.1 (2014): 3307-3329. Available at: http://jco.ascopubs.org/content/32/29/3307.full

Wolff, A.C, Hammond, M.E.H, et.al., "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update." J Clin Onc 31.31 (2013): 3997-4013. Available at: http://jco.ascopubs.org/content/31/31/3997.full

Numerator Statement: HER2-targeted therapies not administered during the initial course of treatment.
Denominator Statement: Adult women with breast cancer that are HER2 negative or HER2 undocumented.
Denominator Exclusions: Patient transfer to practice during or after initial course.

Measure Type: Process
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Oct 22, 2012 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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

- The developer provided evidence that focused on HER2 testing and use of trastuzumab, rather than overuse of trastuzumab.
- The developer provided two clinical practice guidelines:
  - **ASCO Guideline**: ...benefit from adding trastuzumab to chemotherapy was restricted to patients with HER2 protein overexpression and gene amplification (IHC 3+/FISH+), and was not seen in patients with protein overexpession without gene amplifications (IHC 3+/FISH-) or equivocal overexpression with gene amplifications (IHC 2+/FISH+) disease.
  - ...Accuracy of HER2 testing is critical to ensure that those patients most likely to benefit are offered trastuzumab while those unlikely to benefit are spared the cost and toxicity of this agent.
  - ... HER2 As a Predictive Factor for Trastuzumab Benefit in Metastatic Breast Cancer notes that clinical benefit is ´presumed none for patients with negative HER2 testing results.
  - **CCO Guideline**: Trastuzumab should be offered for one year to all patients with HER2-positive node-positive or node-negative, tumour greater than 1 cm in size, and primary breast cancer and who are receiving or have received (neo)adjuvant chemotherapy. Trastuzumab should be offered after chemotherapy.
- The 2012 Steering Committee questioned whether this intervention would happen without HER2 testing or with a negative HER2 result. The developer stated that this can and does happen, according to feedback from payers.

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 updated evidence for this measure was based on 3 clinical practice guidelines:

- American Society of Clinical Oncology (ASCO) guideline on systemic therapy for patients with advanced cancer:
  - Clinicians should recommend HER2-targeted therapy–based combinations for first-line treatment, except for highly selected patients with ER-positive or PgR-positive and HER2-positive disease, for whom clinicians may use endocrine therapy alone. **Strength of recommendation**: Stong. Evidence Quality: High
- Cancer Care Ontario (CCO) guideline on optimal systemic therapy for women with early breast cancer:
  - Only patients with her2-positive breast cancer [ihc 3+, *in situ* hybridization (ish) ratio ≥ 2, or 6+ her2 gene copies per cell nucleus] should be offered adjuvant trastuzumab. **CCO uses a narrative approach to grade the strength of recommendations. No additional details are provided regarding the grading.**
- American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP) joint guideline on HER2

testing:

- Must request HER2 testing on every primary invasive breast cancer (and on metastatic site, if stage IV and if specimen available) from a patient with breast cancer to guide decision to pursue HER2-targeted therapy. This should be especially considered for a patient who previously tested HER2 negative in a primary tumor and presents with disease recurrence with clinical behavior suggestive of HER2-positive or triple-negative disease.
- Must not recommend HER2-targeted therapy if HER2 test result is negative and if there is no apparent histopathologic discordance with HER2 testing (Tables 1 and 2). If the pathologist or oncologist observes an apparent histopathologic discordance after HER2 testing, the need for additional HER testing should be discussed. **Level of evidence**: Not graded

- The developer provided a systematic review of the evidence for the ACSO guideline and included the Quantity, Quality, and Consistency of the evidence.
- The ASCO/CAP joint guideline matches the focus of the measure, however, the level of evidence for this guideline is not graded.

Exception to evidence: Not applicable

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → QQC not provided for ASCO/CAPjoint guideline (Box 4)→ ASCO guideline w/strong level of evidence (Box 5a) → High

Questions for the Committee:

- *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?*

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.

The developer provided the following performance rates from the QOPI® registry:

	2013	2014	2015
# of practices	230	225	265
Total # of patients	6,418	6,168	6,917
Total # patients excluded	122	135	98
Overall	99.39	99.11	99.47
Mean	99.25	99.26	99.54
Min, Max	66.7, 100	84, 100	90.91, 100
Standard Deviation	2.82	2.02	1.38
Percentiles			
P10	100	100	100
P50	100	100	100
P90	96.78	96.67	96.97
P95	95.66	95.66	96.3

- The 2012 Steering Committee expressed concern with the presented performance gap showing concordance of 99 percent with the measure and questioned the opportunity for improvement. The developer stated that the

participants on the measure are a self-selected group participating in the quality Oncology Practice Initiative and performance may be higher for this group. The developer also noted that several unpublished studies suggest overuse of trastuzumab.

Disparities:

- The developer provided the following data on disparities:

	2013	2014	2015
Total # of patients	6,418	6,168	6,917
Overall	99.39	99.11	99.47
Hispanic	100	99.26	99.74
White	99.34	99.20	99.38
Black	98.84	98.47	99.66
Other	98.41	99.43	99.59

Questions for the Committee:

- *What is the quality problem addressed by this measure?*
- *What proportion of patients with breast cancer is represented by this data?*
- This data seems to be capturing a small pecent of the patients – do we expect opportunities for capturing larger number of patients in the measaure?
- *The developer presents some disparities data, are you aware of additional 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 to Support Measure Focus

Comments:

Process measure/group practice level examining the percent of women with HER2 negative (or undocumented) breast cancer who do not received HER2-targeted therapy. Systematic reveiw and guidelines (ASCO--strong/high and CCOntario) demonstrate no clinical benefit from HER2 targeted therapy in HER2 negative breast cancer. HIGH.

1b. Performance Gap

Comments:

QOPI registry data provided for 2013, 2014 and 2015 consistently showing >99% compliance with a narrowing range to 90.9-100% in 2015. Disparity data demonstrate >99% compliance consistently across races with lowest 99.38 for white. It is not clear that there is room for improvement on the current compliance rates for this measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic clinical data: registry

Specifications:

- This is a clinician-level measure.
- The specifications have changed since the last submission

- "Trastuzumab" has been changed to "HER2 targeted therapies" to reflect updated evidence regarding the expansion of treatment options for HER-2 positive patients.

- The numerator of this measure is: HER2-targeted therapies not administered during the initial course of treatment.
- The denominator of this measure is: Adult women with breast cancer that are HER2 negative or HER2 undocumented.
- Patients transferred to practice during or after initial course are excluded from the denominator.
- ICD-10 codes are included; ICD-9 to ICD-10 conversion available on SharePoint.
- The calculation algorithm is provided.
- Instructions for obtaining a minimum sample size are provided.
- The developer specifies how missing data are handled.
- In 2012, the Steering Committee suggested that future iterations of the measure capture:
  - Whether patients are receiving the appropriate dose of hormonal therapy.
  - o the appropriateness of hormonal therapy based upon menopausal state of the patient, and
  - patient adherence to the hormonal therapy through prescription data.

Questions for the Committee :

- *Are the appropriate codes included in the ICD-9 to ICD-10 conversion? Are all appropriate codes included?*
- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 264 patient records from 44 QOPI practices submitted in spring 2007. Trained, independent nurse abstractors from the Virginia Quality Health Center served as the 'gold standard' against which practice abstractions were compared for accuracy.

Describe any updates to testing

- The developer indicated no updates to testing.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Comparing practice abstractions against a 'gold standard' is considered data element validity testing.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing not conducted (Box 2)→Empirical validity testing of patient-level data conducted (Box 3) → Validity testing conducted with patient-level data elements (Box 10) →Appropriate method used but kappa scores for all data elements not provided →Insufficient

Questions for the Committee:

- *See questions under Validity*

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

Rationale: Insufficient information provided about the data elements tested and results.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score was assessed with input from experts involved in ASCO committees in 2011. The developer noted that the face validity survey results revealed that 100% of respondents strongly agreed or agree that this measure provided an accurate reflection of quality and can be used to distinguish good and poor quality. The developer did not provide the number of experts surveyed.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**
☐ **Empirical validity testing of the measure score**

Validity testing method:

- Validity testing was conducted at the patient-level data element for 264 patient records using trained, independent nurse abstractors as the 'gold standard'. Kappa statistics were used to analyze the validity of the audited patient records compared to the submitted patient records. Kappa is the measure of agreement between two raters that adjusts for chance agreements for categorical data. Kappa values range between 0 and 1 and are interpreted as degree of agreement beyond chance. By convention, a kappa > .70 is considered acceptable.

Validity testing results:

- The developer provided a kappa score of 0.74. While this kappa score is above what is considered acceptable, the developer did not state which of the data elements this kappa score represents; no additional results were provided. NQF guidance states that testing should be done for all critical data elements.
- The developer did not state how it was determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.

Questions for the Committee:

- *Does the measure adequately identify and include HER2 negative breast cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on*

validity?Do the results demonstrate sufficient validity so that conclusions about quality can be made?

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients transferred to practice during or after initial course are excluded from the denominator.
- The developer did not provide an analysis of the exclusions.

Questions for the Committee:

- *Without an analysis of the exclusions, can the Committee determine:*
  - *if any patients or patient groups inappropriately excluded from the measure?*
  - *the exclusions are of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reported:

- Data from thefall 2011 QOPI round, data was submitted October and November 2011. 204 practices reported this measure. Data from 5,968 patient records were submitted for this measure.

Year	Mean	St. Dev.	Min	10th	25th	50th	75th	90th	Max
2013	99.25	2.82	66.7	100	100	100	100	96.78	100
2014	99.26	2.02	84	100	100	100	100	96.67	100
2015	99.54	1.38	90.91	100	100	100	100	96.97	100

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

The developer indicates N/A.

2b7. Missing Data

The developer provided the following information:

- This measure is specified with defined criteria and data elements. If a patient record does not include one or more of these components for the initial patient population or denominator, then patients are not considered eligible for the measure and not included.
- If data to determine whether a patient should be considered for the numerator or exclusions is missing, then the numerator or exclusions not considered to be met and the practice will not get credit for meeting performance for that patient.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly addressed (Box2) → Empirical validity testing conducted (Box 3) →Validity testing not conducted at the measure score (Box 6)→Validity testing conducted with patient-level data elements (Box 10) →Appropriate method used but kappa scores for all data elements not provided →Insufficient

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

Rationale: Insufficient information provided about the data elements tested and results, but if all data elements had similar Kappa scores, then the rating would be moderate.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
Numerator: HER2-targeted therapy (previously "trastuzumab") not administered. Demonimator: adult women with HER2 negative (or undocumented) breast cancer. Definitions are clear and appropriate codes captured. No concerns about implementation.
Consistent with evidence.

2a2. Reliability Testing
Comments:
No updates to original data element testing that used 44 QOPI practices compared to Virginia Health Center.

2b2. Validity Testing
Comments:
Face validity testing was reportedly done. Kappa score 0.74 cited without specifics.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Stated exclusions include care by prior providers. no provision is made for reveiw of exclusions so cannot evaluate impact on 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.

- All data elements are in defined fields in electronic claims.
- Data elements are generated or collected by and used by healthcare personnel during the provision of care.
- In 2012, the Steering Committee raised concerns that extraction of this data may be burdensome as it may require chart abstractions. Eventual use of this measure through EHRs would lessen this burden.

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
All defined electronic claims fields reliably collected. Feasible with EHR or chart abstraction.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- The measure is used in Quality Oncology Practice Initiative (QOPI®).
- The measure was recently selected for inclusion in a Medical Oncology Core Measure Set supported by America's Health Insurance Plans and CMS.

Improvement results

- The developer noted performance continues to remain high, some variation remains from year to year.
- Currently, performance results are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program. The developer is hopeful additional performance rates across the U.S. will become available with the AHIP Core Measures Collabortiave.

Unexpected findings (positive or negative) during implementation: The developer reported no challenges or unexpected findings in implementation.

Potential harms : The developer reported no unintended consequences were noted during testing.

Feedback :

- MAP has not reviewed this measure for inclusion in any federal program.
- During the endorsement process in 2012, the Steering Committee, NQF members, and the public provided the following comments:
    - The Steering Committee members expressed concern that several measures had high rates of performance, indicating a small gap in performance; however, the developer clarified that the performance gap data came from the American Society for Clinical Oncology's Quality Oncology Practice Initiative (QOPI), which included self-selecting practices voluntarily reporting on measures. As such, the developer stated that it is likely that there is more variation in performance than was demonstrated through QOPI. The Steering Committee agreed with the developer that it is likely that there is variation in use of trastuzumab and in HER2 testing, given the self-selecting nature of the practices participating with QOPI. Taken in conjunction with several studies suggesting overuse of trastuzumab, the Steering Committee recommended the measure for endorsement.
    - Public and member comments included:
        - A recommendation that references to the specific therapy, trastuzumab, be changed to "FDA-approved HER2 therapy."
        - A recommendation against endorsement of the measure due to limited utility in improving quality, citing a 2009 study where 98 percent of patients had HER2 testing and 100 percent of patients receiving trastuzumab had documented HER2 testing prior to receiving trastuzumab.
        - A recommendation that a HER2 composite measure be developed, comprised of measures 1857, 1855, 1858, and 1878.
    - The developer responded:
        - The preponderance of available data suggest room for improvement. The developer noted that oncologists need to know the result of HER2 testing that was accomplished prior to oncologist engagement. HER2 status should be captured in a way that can be located/retrieved from the medical record. The developer stated that given the large numbers of women affected, modest improvements can have a significant national impact. Lastly, the developer noted that if ongoing use of this measure - or the underuse of trastuzumab measure - reveals in future years that no quality gap exists, ASCO will retire the measure.
        - The developer stated that ASCO and CAP, the developers of the referenced measures, have discussed the concept of a composite measure, and neither organization believes that it is advantageous at this time. These measures are designed for different providers and levels of accountability, and have different denominators. Measure 1855 was developed to measure the

performance of individual pathologists, while measures 1857, 1858, and 1878 are for medical oncologists/clinical oncology practices. It may be beneficial to implement all of these measures within certain settings, such as accountable care organizations or Cancer Care Centers. ASCO reports measures 1857, 1858, and 1878 together in their quality programs; however, they believe that the measures are independently useful. The developer will consider paired or composite measures in the future. The Steering Committee agreed that as the measures are currently specified for different levels of analysis, a composite measure would not be feasible. Further, the Steering Committee agreed that the measures capture discrete steps in care.

***Questions for the Committee**:*

- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Although the measure is in use, is the small (to no) improvement over time indicative of poor usability?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Currenly used in accountability program (QOPI and Core CMS). As current data are based on voluntary (QOPI) reporting, developers suggest that the gap will be more apparent with broader, mandated application. ?consolidation. No unintended consequences.

Criterion 5: Related and Competing Measures

Related or competing measures

- No related or competing measures have been identified by the developer.

Pre-meeting public and member comments

-

PREVIOUS EVIDENCE FORM

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1857 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
This process measure assesses overuse of trastuzumab in women with HER2/neu-negative breast cancer. It is critical to ensure that patients unlikely to benefit are spared the cost and toxicity of this agent.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The guidelines and RCTs referenced focused on HER2 testing and use of trastuzumab, rather than oversue of trastuzumab.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: Six randomized control trials were considered by Cancer Care Ontario. The NCCN guideline considered five clinical trials. The ASCO/CAP guideline included five clinical trials.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The guidelines and RCTs referenced focused on HER2 testing and use of trastuzumab, rather than overuse of trastuzumab. The evidence described in the RCTs relates to use of tratuzumab in patients with HER2 positive breast cancer. Results from these studies reveal that the benefit of trastuzumab in women with HER2 negative breast cancer is "presumed none." One trial specifically considered cardiac adverse events to assess potential harms, other trials considered adverse events in addition to disease-specific outcomes.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* n/a The guidelines and RCTs referenced focused on HER2 testing and use of trastuzumab, rather than overuse of trastuzumab.

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
The HER2 gene is amplified in approximately 18% to 20% of breast cancers. There are no known benefits of trastuzumab in breast cancer patients who are HER2 negative.

Trastuzumab therapy is not without its drawbacks. Although treatment duration in the metastatic setting varies widely, currently adjuvant trastuzumab is recommended for 12 months. The drug cost of 52 weeks of trastuzumab in the community setting in the United States is approximately $100,000 based on average sales price (www.accc-cancer.org). In addition, there is a requirement for 9 to 12

months of intravenous therapy after completion of adjuvant chemotherapy.

Importantly, trastuzumab is associated with a small risk of serious cardiac toxicity. In the prospective randomized adjuvant trials, careful serial cardiac monitoring has demonstrated that at median follow-up times of 3 years or fewer, approximately 5% to 15% of patients develop cardiac dysfunction, and approximately 1% to 4% develop significant cardiac events (including symptomatic congestive heart failure) while taking trastuzumab. Taken together, the high cost and potential cardiotoxicity demand appropriate use of trastuzumab.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** CCO Guidelines use a narrative approach in grading the quality of the evidence.

NCCN also grades the body of evidence used to develop recommendations.

The panel that develops guideline recommendations for NCCN grades the recommendations. Participants in the panel primarily include subject matter experts. They do not require disclosures about potential conflicts of interest.

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** CCO Guidelines use a narrative approach in grading the quality of the evidence.

NCCN grades the body of evidence according to a system developed by that organization, "NCCN Categories of Evidence and Consensus".

1c.13 **Grade Assigned to the Body of Evidence:**

1c.14 **Summary of Controversy/Contradictory Evidence:** No known evidence that contradicts current recommendations was noted.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

1c.16 **Quote verbatim, <u>the specific guideline recommendation</u>** *(Including guideline # and/or page #)*:
ASCO Guideline: ...benefit from adding trastuzumab to chemotherapy was restricted to patients with HER2 protein overexpression and gene amplification (IHC 3+/FISH+), and was not seen in patients with protein overexpession without gene amplifications (IHC 3+/FISH-) or equivocal overexpression with gene amplifications (IHC 2+/FISH+) disease.(Page 138).
...Accuracy of HER2 testing is critical to ensure that those patients most likely to benefit are offered trastuzumab while those unlikely to benefit are spared the cost and toxicity of this agent. (Page 138)
... Table A7. HER2 As a Predictive Factor for Trastuzumab Benefit in Metastatic Breast Cancer (Page 138) notes that clinical benefit is ´presumed none for patients with negative HER2 testing results.

CCO Guideline: Trastuzumab should be offered for one year to all patients with HER2-positive node-positive or node-negative, tumour greater than 1 cm in size, and primary breast cancer and who are receiving or have received (neo)adjuvant chemotherapy. Trastuzumab should be offered after chemotherapy. (CCO guideline, development and methods pg 3/ pdf pg 29; https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=13890).

1c.17 **Clinical Practice Guideline Citation:** Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, van de Vijver M, Wheeler TM, Hayes DF; American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncolol. 2007;25:118-145.

Members of the Breast Cancer Disease Site Group. The role of trastuzumab in adjuvant and neoadjuvant therapy in women with HER2/neu-overexpressing breast cancer. Madarnas Y, Tey R, reviewers. Toronto (ON): Cancer Care Ontario; 2011 Sep 15 [Endorsed 2010 Jun 11]. Program in Evidence-based Care Evidence-Based Series No.: 1-24 Version 2.

NCCN Clinical Practice Guidelines in Oncology. Breast Cancer, Version 2.2011. NCCN.org.

1c.18 **National Guideline Clearinghouse or other URL:** https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=13890 AND http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf AND http://jco.ascopubs.org/content/25/1/118.pdf

1c.19 **Grading of Strength of Guideline Recommendation.** Has the recommendation been graded? Yes

1c.20 **If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** Recommendations are graded by the panel that creates those recommendations. Disclosures are required to minimize potential bias. The panel that develops guideline recommendations for NCCN grades the recommendations. Participants in the panel primarily include subject matter experts.

1c.21 **System Used for Grading the Strength of Guideline Recommendation:** Other

1c.22 **If other, identify and describe the grading scale with definitions:** CCO Guidelines use a narrative approach to grade the strength of recommendations.

NCCN grades recommendations depending upon the strength, directness, precision, and other factors related to the evidence underlying the recommendation.

1c.23 **Grade Assigned to the Recommendation:**

1c.24 **Rationale for Using this Guideline Over Others:**

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

UPDATED EVIDENCE FORM

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1857

Measure Title: HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies

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: 3/11/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

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

Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

☐ Health 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

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

☒ Process: HER2 targeted therapy spared for patients who are HER2 negative or undocumented

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

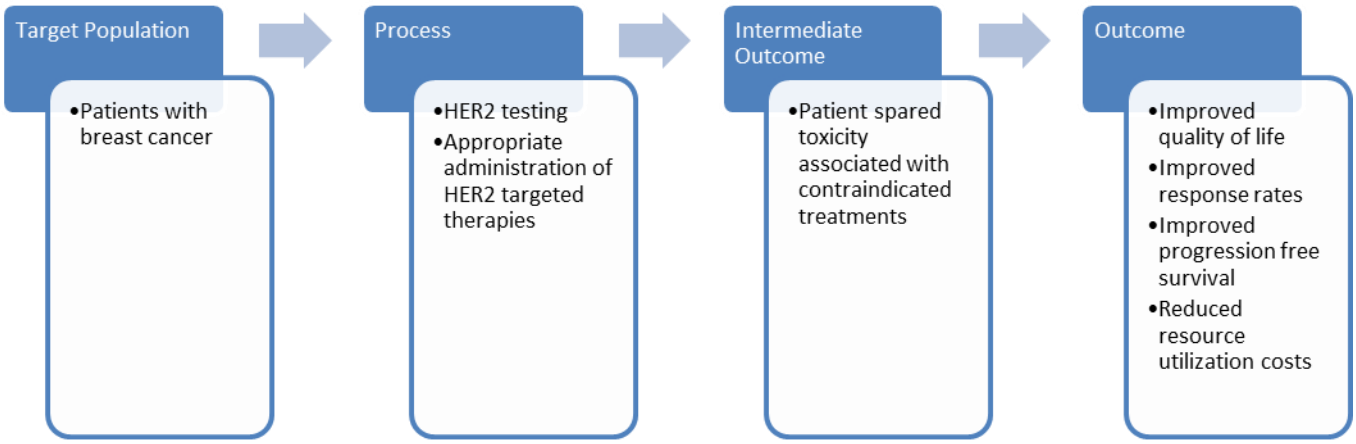
1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

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

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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



Human epidermal growth factor receptor (HER2) gene is amplified and/or overexpressed in approximately 15% to 20% of primary breast cancers (Giordano, 2014). The ASCO/CAP joint guideline on HER2 testing recommends all patients with invasive breast cancer should be tested for HER2 status and only those who test positive for HER2 status should receive HER2 targeted therapies. Additionally data have shown that the administration of HER2 targeted therapies such as Pertuzumab offer no clinical benefit in patients with HER2 negative metastatic disease (Wolff, 2013).

The contraindicated administration of HER2 targeted therapy to patients with HER2 negative breast cancer can propagate potentially toxic, costly and adverse effects as well as decrease the patient's overall quality of life (Partridge, 2014).

Citations:

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:

http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Partridge, A.H., Smith, I.E., et. al., "Chemo- and Targeted Therapy for Women with Human Epidermal Growth Factor Receptor 2- Negative (or Unknown) Advanced Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Onc Pr 11.1 (2014): 3307-3329. Available at:
http://jco.ascopubs.org/content/32/29/3307.full

Wolff, A.C, Hammond, M.E.H, et.al., "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update." J Clin Onc 31.31 (2013): 3997-4013. Available at:
http://jco.ascopubs.org/content/31/31/3997.full

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

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

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

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

☐ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:

http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Eisen, A., K.G, Fletcher, et.al, "Optimal Systemic Therapy for Early Breast Cancer in Women: A Clinical Practice Guideline." Curr Onc 22.0 (2014): Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4381792/

Wolff, A.C, Hammond, M.E.H, et.al., "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update." J Clin Onc 31.31 (2013): 3997-4013. Available at: http://jco.ascopubs.org/content/31/31/3997.full

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

ASCO guideline on systemic therapy for patients with advanced cancer:

Pg. 2081

Recommendation 1:

"Clinicians should recommend HER2-targeted therapy–based combinations for first-line treatment, except for highly selected patients with ER-positive or PgR-positive and HER2-positive disease, for whom clinicians may use endocrine therapy alone." Strength of recommendation: Strong. Evidence Quality: High

CCO guideline on optimal systemic therapy for women with early breast cancer:

Page S75

Recommendation 26:

"Only patients with her2-positive breast cancer [ihc 3+, *in situ* hybridization (ish) ratio ≥ 2, or 6+ her2 gene copies per cell nucleus] should be offered adjuvant trastuzumab."

ASCO/CAP Joint Guideline on HER2 Testing:

Page 3998:

"Must request HER2 testing on every primary invasive breast cancer (and on metastatic site, if stage IV and if specimen available) from a patient with breast cancer to guide decision to pursue HER2-targeted therapy. This should be especially considered for a patient who previously tested HER2 negative in a primary tumor and presents with disease recurrence with clinical behavior suggestive of HER2-positive or triple-negative disease"

"Must not recommend HER2-targeted therapy if HER2 test result is negative and if there is no apparent histopathologic discordance with HER2 testing (Tables 1 and 2). If the pathologist or oncologist observes an apparent histopathologic discordance after HER2 testing, the need for additional HER2 testing should be discussed."

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

ASCO guideline: Strength of recommendation: Strong.

Definition: There is high confidence that the recommendation reflects best practice. This is based on (1) strong evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with no or minor exceptions; (3) minor or no concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation

CCO guideline: CCO Guidelines use a narrative approach to grade the strength of recommendations. Additional details are not provided in the guideline.

ASCO/CAP Joint Guideline: recommendation not graded.

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

ASCO Guideline:

Rating for Strength of Recommendation	Definition
Strong	There is high confidence that the recommendation reflects best practice. This is based on (1) strong evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with no or minor exceptions; (3) minor or no concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation.
Moderate	There is moderate confidence that the recommendation reflects best practice. This is based on (1) good evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with minor and/or few exceptions; (3) minor and/or few concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a moderate recommendation.
Weak	There is some confidence that the recommendation offers the best current guidance for practice. This is based on (1) limited evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, but with important exceptions; (3) concerns about study quality; and/or (4) the extent of panelists' agreement. Other considerations (discussed in the guideline's literature review and analyses) may also warrant a weak recommendation.

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

http://www.instituteforquality.org/sites/instituteforquality.org/files/her2_treatment_ms_5.21.pdf

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

☒ Yes → ***complete section 1a.7***

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

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

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

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

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

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

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

Complete section 1a.7

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

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

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

Complete section 1a.7

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

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency

of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

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

ASCO Guideline:

p. 2078

To provide evidence-based recommendations to practicing oncologists and others on systemic therapy for patients with human epidermal growth factor receptor 2 (HER2) –positive advanced breast cancer.

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

Evidence Quality: High

Definition: High confidence that the available evidence reflects the true magnitude and direction of the net effect (i.e. balance of benefits v harms) and that further research is very unlikely to change either the magnitude or direction of this net effect.

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

Rating for Strength of Evidence	Definition
High	High confidence that the available evidence reflects the true magnitude and direction of the net effect (i.e., balance of benefits *v* harms) and that further research is very unlikely to change either the magnitude or direction of this net effect.
Intermediate	Moderate confidence that the available evidence reflects the true magnitude and direction of the net effect. Further research is unlikely to alter the direction of the net effect; however, it might alter the magnitude of the net effect.
Low	Low confidence that the available evidence reflects the true magnitude and direction of the net effect. Further research may change either the magnitude and/or direction this net effect.
Insufficient	Evidence is insufficient to discern the true magnitude and direction of the net effect. Further research may better inform the topic. The use of the consensus opinion of experts is reasonable to inform outcomes related to the topic.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1966-2012

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

11 randomized controlled clinical trials

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ASCO guideline:

P. 2081

This recommendation is based on a body of evidence regarding first-line therapy, found both in the ASCO and CCO systematic reviews. CCO included the pivotal trial by Slamon et al and nine other RCTs of trastuzumab. These trials found a benefit for HER2-targeted therapy combinations, specifically with trastuzumab. The study by Slamon et al was the only first-line phase III trial that compared an HER2-targeted therapy plus chemotherapy with chemotherapy alone. That trial found survival, time to progression (TTP), and overall response rate benefits in the trastuzumab arm. The CCO review found two phase III trials that compared HER2-targeted therapy plus endocrine therapy with endocrine therapy alone. Both of those trials found progression-free survival (PFS) and TTP benefits, but no overall survival (OS) benefit, in the combination arm and will be discussed in the section on endocrine therapy (Clinical Question 2), along with another more recent trial.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

ASCO guideline:

P. 2092

Overall, HER2-targeted therapy in combination with chemotherapy in the first-line setting is associated with improvements in response rate, PFS (progression-free survival), TTP (time to progression), and OS (overall survival) when compared with chemotherapy alone. In trials of endocrine therapy, the addition of HER2-targeted therapy is associated with improvements in response rate and PFS but not in survival. These data support the use of HER2-targeted therapy in the first-line treatment of metastatic breast cancer.

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

ASCO guideline:

P.2092

There are some contraindications to HER2-targeted therapy, as a result of its cardiovascular toxicity effects (Table 4). The single most important contraindication is a decreased left ventricular ejection fraction (LVEF) and/or clinical evidence of congestive heart failure arising from low LVEF. Among patients with congestive heart failure or low ejection fraction, the decision to use HER2-targeted therapy must be made on an individual basis, assessing the relative risks of cardiac dysfunction from a specific regimen versus disease progression.

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

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

No relevant studies have been conducted and published since the systematic reviews.

1a.8 OTHER SOURCE OF EVIDENCE

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

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

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

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
BREAST_1857_MeasSubm_Evidence_2013-08-20-635933068979914314.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
Human epidermal growth factor receptor (HER2) gene is amplified and/or overexpressed in approximately 15% to 20% of primary breast cancers (Giordano, 2014). The ASCO/CAP joint guideline on HER2 testing recommends all patients with invasive breast cancer should be tested for HER2 status and only those who test positive for HER2 status should receive HER2 targeted therapies. Additionally data have shown that the administration of HER2 targeted therapies such as Pertuzumab offer no clinical benefit in patients with HER2 negative metastatic disease (Wolff, 2013).

The contraindicated administration of HER2 targeted therapy to patients with HER2 negative breast cancer can propagate potentially toxic, costly and adverse effects as well as decrease the patient's overall quality of life (Partridge, 2014).

Citations:

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Partridge, A.H., Smith, I.E., et. al., "Chemo- and Targeted Therapy for Women with Human Epidermal Growth Factor Receptor 2-Negative (or Unknown) Advanced Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Onc Pr 11.1 (2014): 3307-3329. Available at: http://jco.ascopubs.org/content/32/29/3307.full

Wolff, A.C, Hammond, M.E.H, et.al., "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update." J Clin Onc 31.31 (2013): 3997-4013. Available at: http://jco.ascopubs.org/content/31/31/3997.full

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This data was produced from the QOPI® registry and data was abstracted for a sample of patients seen with the data collection period. Performance is reported at the clinical practice level.

In 2013, 230 practices were measured, and the total patient population for this measure was 6418. 122 total patients were excluded across all reporting practices.

In 2014, 225 practices were measured, and the total patient population for this measure was 6168. 135 total patients were excluded across all reporting practices.

In 2015, 265 practices were measured, and the total patient population for this measure was 6917. 98 total patients were excluded across all reporting practices.

_____2013 2014 2015

Overall	99.39	99.11	99.47
Mean	99.25	99.26	99.54
Minimum	66.7	84	90.91
Maximum	100	100	100
Standard Deviation	2.82	2.02	1.38
Percentiles			
P10	100	100	100
P25	100	100	100
P50	100	100	100
P75	100	100	100
P90	96.78	96.67	96.97
P95	95.66	95.66	96.3

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.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
This data was produced from the QOPI® registry and data was abstracted for a sample of patients seen with the data collection period. Performance is reported at the chart level.

In 2013, the total patient population for this measure was 6418.

In 2014, the total patient population for this measure was 6168.

In 2015, the total patient population for this measure was 6917.

_____	2013	2014	2015
Overall	99.39	99.11	99.47
Hispanic	100	99.26	99.74
White	99.34	99.20	99.38
Black	98.84	98.47	99.66
Other	98.41	99.43	99.59

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Health disparities between patients with breast cancer according to race/ethnicity, age, insurance status, geographic location, education, and other factors are well documented, however literature addressing disparities specific to patients with HER2-positive metastatic breast cancer is scarce. According to some studies, there are not large (although some suggest modest) differences in the prevalence of HER2 positivity between women with breast cancer of different races/ethnicities. The variation by race is smaller among those with HER2-positive breast cancer than for some other subtypes.

HER2 positivity is not necessarily associated with worse treatment outcomes among African American compared with non–African American patients. However, high-quality data on patients with HER2-positive metastatic disease are still needed to reach conclusions related to outcomes based on ethnicity. Therefore, health disparities may be similar to those faced by patients with metastatic breast cancer generally.

Although ASCO clinical practice guidelines represent expert recommendations the highest level of cancer care, it is important to note that many patients have limited access to medical care. Racial and ethnic disparities in health care contribute significantly to this problem in the United States. Minority racial/ethnic patients with cancer suffer disproportionately from comorbidities, experience more substantial obstacles to receiving care, are more likely to be uninsured, and are at greater risk of receiving care of poor quality than other North Americans. Many other patients lack access to care because of their age, geography, and distance from appropriate treatment facilities. Awareness of these disparities in access to care should be considered in the context of this clinical practice

guideline, and health care providers should strive to deliver the highest level of cancer care to these vulnerable populations.

Citations:

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Breast

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse

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.)*
No webpage available

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.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
"Trastuzumab" has been changed to "HER2 targeted therapies" to reflect updated evidence regarding the expansion of treatment options for HER-2 positive patients.

Changes to the measure were made after the latest measure update of ASCO's Quality Oncology Practice Initiative (QOPI®) measures and therefore the data and testing reflect the previous version of the measure. These changes will be implemented in the Fall of 2016.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
HER2-targeted therapies not administered during the initial course of treatment.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The initial course of treatment is defined as: The treatment course for the initial diagnosis, which may include elements of chemotherapy (any route), hormonal therapy, radiation, or additional surgery. Do not include treatment provided for recurrence or disease progression.

S.6. **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, 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.
HER2 targeted therapy administered during initial treatment course = HER2 targeted therapy NOT administered
OR

HER2 targeted therapy administered during initial treatment course = HER2 targeted therapy administered
AND
HER2 targeted therapy administered according to clinical trial protocol = Yes)

'HER2 targeted therapies' include trastuzumab, pertuzumab, T-DM1.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Adult women with breast cancer that are HER2 negative or HER2 undocumented.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Female
And
2 or more encounters at the reporting site
And
Age at diagnosis greater than or equal to 18 years
And
Initial breast cancer diagnosis [C50.01-, C50.11-, C50.21-, C50.31-, C50.41-, C50.51-, C50.61-, C50.81-, C50.91-]
AND
(HER-2/neu status = HER2 negative
OR
HER-2/neu status = Test ordered, results not yet documented
OR
HER-2/neu status = Test NOT ordered/no documentation
OR
HER-2/neu status=Test ordered, insufficient sample for results
Or
HER-2/neu status= HER2 equivocal)

Definitions
Encounter: Patients must have been first seen in the office by a medical oncology or hematology oncology practitioner for the cancer diagnosis eligible for inclusion within the 1-year time frame of the reporting period. Enter the most recent visit that occurred during the 6-month visit window before the abstraction date. This can include visits to other office sites within the practice only if the practice uses a common medical record and shares management of care for the patient. This does not include visits during which a practitioner wasn´t seen (e.g., laboratory testing), inpatient consults/visits, phone or email consults, or visits to a surgeon or radiation oncologist.

HER2 status:

Select 'Test ordered, results not yet documented´ only if there is documentation in the chart that a test that included HER2 analyses was ordered.
In the absence of any documentation regarding HER-2/neu status, select 'Test not ordered/no documentation.'

Enter information from the most recent test report. If the most recent report indicates insufficient sample, select 'Test ordered, insufficient sample for results.'
If a physician note and the HER-2/neu report differ in results, report the status in the physician note if the note explains the discrepancy. Otherwise, report the status from the HER-2/neu report.

Use the following definitions to determine HER-2/neu status:

Positive:
IHC 3+ based on circumferential membrane staining that is complete, intense

- ISH positive based on:
- Single-probe average HER2 copy number =6.0 signals/cell
- Dual-probe HER2/CEP17 ratio =2.0 with an average HER2 copy number =4.0 signals/cell
- Dual-probe HER2/CEP17 ratio =2.0 with an average HER2 copy number <4.0 signals/cell
- Dual-probe HER2/CEP17 ratio < 2.0 with an average HER2 copy number =6.0 signals/cell

Equivocal:
- IHC 2+ based on circumferential membrane staining that is incomplete and/or weak/moderate and within > 10% of the invasive tumor cells or complete and circumferential membrane staining that is intense and within = 10% of the invasive tumor cells
ISH equivocal based on:
- Single-probe ISH average HER2 copy number = 4.0 and < 6.0 signals/cell
- Dual-probe HER2/CEP17 ratio < 2.0 with an average HER2 copy number = 4.0 and < 6.0 signals/cell

Negative:
IHC 1+ as defined by incomplete membrane staining that is faint/barely perceptible and within > 10% of the invasive tumor cells or IHC 0 as defined by no staining observed or membrane staining that is incomplete and is faint/barely perceptible and within = 10% of the invasive tumor cells
ISH negative based on:
- Single-probe average HER2 copy number < 4.0 signals/cell
- Dual-probe HER2/CEP17 ratio < 2.0 with an average HER2 copy number < 4.0 signals/cell

Indeterminate:
Indeterminate if technical issues prevent one or both tests (IHC and ISH) from being reported as positive, negative, or equivocal.
Conditions may include:
- Inadequate specimen handling,
- Artifacts (crush or edge artifacts) that make interpretation difficult
- Analytic testing failure.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patient transfer to practice during or after initial course.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
Transfer-in Status does not equal Reporting practice has/had primary responsibility for the initial course of the patient´s medical oncology care

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not applicable

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
Performance is calculated as:

1. Identify those patients that meet the denominator criteria defined in the measure.

2. Subtract those patients with a denominator exclusion from the denominator if applicable.

3. From the patients who qualify for the denominator (after any exclusions are removed), identify those who meet the numerator criteria.

4. Calculation: Numerator/Denominator-Denominator Exclusions

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
QOPI registry abstraction is offered twice a year to participating Medical Oncology and Hematology Oncology Practices. The minimum sample size for each data abstraction period is based on the number of med-onc and hem-onc FTEs at the practice and/or site level. For breast cancer, patients must be female, 18 years and older with a diagnosis of [C50.01-, C50.11-, C50.21-, C50.31-, C50.41-, C50.51-, C50.61-, C50.81-, C50.91-] within the one year diagnosis window applicable to the round. The practices follow a chart selection methodology which identify patients who had a diagnosis date within one year of the abstraction period start date AND had two office visits with a practitioner in the office within three months of the abstraction data period start date.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
This measure is specified with specific criteria and data elements. If a patient record does not include one or more of these components for the initial patient population or denominator, then patients are not considered eligible for the measure and not included.

If data to determine whether a patient should be considered for the numerator or exclusions is missing, then the numerator or exclusions not considered to be met and the practice will not get credit for meeting performance for that patient.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
ASCO Quality Oncology Practice Initiative (QOPI®)

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1857_MeasureTesting_Data_1857_Update.doc

NQF #: 1857 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
2008 IRR study: A sample of 300 records was planned for re-abstraction in four geographic regions: Midwest, Northeast, South, and West. 50 QOPI practices were randomly selected from the 4 geographic areas and invited to participate. Within each practice, six previously abstracted charts were selected randomly for re-abstraction from the population of 13,561 records submitted in spring 2007 round. Forty-four practices agreed to participate, and submitted 264 records (6 per practice).

2010-2011 audit: QOPI practices applying for the QOPI Certification Program are required to submit copies of documentation from 3-5 records which were previously abstracted. Trained ASCO auditors randomly select records within each domain for audit. Agreement at the data element level is documented. 426 audited records from 130 practices were complete in November 2011 and included in the concordance analysis.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**
2008 IRR study: ASCO engaged the Virginia Quality Health Center to conduct an inter-rater reliability study of the QOPI case report form and measures. Trained, independent nurse abstractors served as the 'gold standard' against which practice abstractions were compared for accuracy. Sampling is described above. The 264 sampled records allowed for reliability analysis at a 95% confidence level with a +/- 3.88% marking of error.
Kappa statistics were used to analyze the reliability of the audit data set compared to the submitted data. Kappa statistics are the commonly accepted standard for determining inter-rater reliability in the healthcare setting (Allison, Calhoun, et al, 2000; Cassidy, Marsh, et al, 2002). The Kappa statistic is conceptually similar to the rate of agreement between two reviewers, but it imposes a more stringent standard that simple agreement and mismatch rates. The following standards were used (Cohen, 1960; Sim and Wright, 2005; Feinstein and Cicchetti, 1990):

- Kappa > .0.75 denotes excellent reliability
- Kappa between 0.40 and 0.75 denotes good reliability
- Kappa less than 0.40 denote marginal reliability

2010-2011 audit: Agreement data from 426 records were imported into a formatted data table for analysis. First, agreement data were used to calculate concordance at the data element level. Second, by applying the measure analytic calculation, concordance at the measure level was calculated.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
2008 IRR study: measure level Kappa 0.74 (good reliability). Specifications and instructions were updated based on results

2010-2011 audit: measure level concordance 96% (valid N=133 records)

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
Measure specifications are consistent with the evidence cited.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
In 2009, an ASCO steering group comprised of medical oncologists, health services researchers, and quality experts undertook an iterative, criteria-based assessment process to identify QOPI measures that are appropriate for use for accountability measurement. This measure was selected as appropriate for accountability.

Face validity of the measure score was assessed via survey of experts involved in ASCO committees in 2011. The survey explicitly asked whether the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
Face validity survey results revealed that 100% of respondents 'strongly agree' or 'agree' that this measure provides an accurate reflection of quality and can be used to distinguish good and poor quality.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
n/a

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
n/a

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
n/a

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
n/a

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
n/a

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
n/a

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
n/a

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Data reported are from the Fall 2011 QOPI round, reflecting data submitted October and November 2011. 204 practices reported this measure. Data from 5968 patient records were submitted for this measure.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
QOPI measure analytics at the practice level were generated. Practices with fewer than 5 records were not included in calculations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

Year	Mean	St. Dev.	Min	10th	25th	50th	75th	90th	Max
2013	99.25	2.82	66.7	100	100	100	100	96.78	100
2014	99.26	2.02	84	100	100	100	100	96.67	100
2015	99.54	1.38	90.91	100	100	100	100	96.97	100

This measure has been implemented in QOPI for several years. In this self-selected group on oncology practitioners committed to quality assessment and improvement, concordance with this measure has been high overall; however, a proportion of practices (new and experienced) continue to demonstrate sub-optimal variation.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
n/a

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
n/a

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
n/a

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: n/a

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
n/a

2.1-2.3 Supplemental Testing Methodology Information:
Attachment
1c.8_response_for_1857_trastuzumab_not_adminstered.docx

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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)*

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

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure and its specifications have been in place for several years and ASCO continues to monitor and ensure that the measure and its specifications are up to date for widespread use.

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

Not applicable

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
Payment Program	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/manual-qopi-measures

4a.1. **For each CURRENT use, checked above, provide:**

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

Quality Oncology Practice Initiative:

In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

4a.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?*)

We are continuously seeking opportunities to advocate for expanded use of this measure in government or other programs, including those intended for accountability or public reporting. For example, this measure was recently selected for inclusion in a Medical Oncology Core Measure Set supported by America's Health Insurance Plans and CMS. See section 4a.3. below for additional details.

4a.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 has also been included in America's Health Insurance Plans Medical Oncology Core Measure Set. The purpose of this program is to reduce variability in measure selection, specifications and implementation. The measures will be implemented nationally by private health plans using a phased-in approach. Contracts between physicians and private payers are individually negotiated and therefore come up for renewal at different points in time depending on the duration of the contract. It is anticipated that private payers will implement these core sets of measures as and when contracts come up for renewal or if existing contracts allow modification of the performance measure set. CMS is also working to align measures across public programs. They intend to include, for broad input, the agreed upon draft measure sets in the Physician Fee Schedule and other proposed rules. For measures that are not currently in CMS programs, CMS would go through the annual pre-rulemaking and rulemaking processes to solicit stakeholder and public input. Depending on public response, these measures will be included in a timeframe determined by the Agency.

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

While performance continues to be generally high, some variation still remains as evidenced by the performance ranges by year and from the time of the last NQF endorsement review. The data available are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program.

Additional information on overall performance rates across the U.S. will hopefully become available with the AHIP Core Measures Collaborative.

4b.2. **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.**
While performance continues to be generally high, some variation still remains as evidenced by the performance ranges by year and from the time of the last NQF endorsement review. The data available are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program.

Additional information on overall performance rates across the U.S. will hopefully become available with the AHIP Core Measures Collaborative.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There have been no reports of unintended consequences with 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.
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

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 completely harmonized?

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 **Attachment:** QOPI_Adoption_of_ICD10_020916-635933001750874650.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Society of Clinical Oncology
Co.2 **Point of Contact:** Tayyaba, Shehzadi, Tayyaba.Shehzadi@asco.org, 571-483-1673-
Co.3 **Measure Developer if different from Measure Steward:** American Society of Clinical Oncology
Co.4 **Point of Contact:** Tayyaba, Shehzadi, Tayyaba.Shehzadi@asco.org, 571-483-1673-

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.
ASCO Breast Cancer Measures Development Panel
The panel is responsible for reviewing evidence and maintaining measures

Gary Lyman, MD, MPH, FASCO, FRCP
Co-Chair
Fred Hutchinson Cancer Research Center

Gabrielle Rocque, MD
Co-Chair
University of Alabama

Banu Arun, MD
University of Texas
MD Anderson Cancer Center

Gary Cohen, MD, FASCO
Cancer Center at GBMC

Shelley Fuld Nasso, MPP
National Coalition for Cancer Survivorship

Jennifer Griggs, MD, MPH
University of Michigan

Michael Hassett, MD, MPH
Dana-Farber Cancer Institute

Michael Neuss, MD, FASCO
Vanderbilt Ingram Cancer Center

Ann Partridge, MD, FASCO
Dana-Farber Cancer Institute

Michael Soble, MD
North Shore Oncology/Hematology Associates

Ann Von Gehr, MD
Permanente Medical Group Inc.

Antonio Wolff, MD, FACP, FASCO
Johns Hopkins Kimmel Cancer Center

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 02, 2016
Ad.4 **What is your frequency for review/update of this measure?** q3years
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2017

Ad.6 **Copyright statement:** The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for

commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] or American Society of Clinical Oncology (ASCO). Neither the AMA, ASCO, PCPI, nor its members shall be responsible for any use of the Measures.

The AMA's and PCPI's significant past efforts and contributions to the development and updating of the Measures is acknowledged. ASCO is solely responsible for the review and enhancement ("Maintenance") of the Measures as of 2016.

ASCO encourages use of the Measures by other health care professionals, where appropriate.

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

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. The AMA, ASCO, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments:

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 #: 1878
Measure Title: HER2 testing for overexpression or gene amplification in patients with breast cancer
Measure Steward: American Society of Clinical Oncology
Brief Description of Measure: Proportion of female patients (aged 18 years and older) with breast cancer who receive human epidermal growth factor receptor 2 (HER2) testing for overexpression or gene amplification
Developer Rationale: Human epidermal growth factor receptor (HER2) gene is amplified and/or overexpressed in approximately 15% to 20% of primary breast cancers. The ASCO/CAP joint guideline on HER2 testing recommends all patients with invasive breast cancer should be tested for HER2 status and only those who test positive for HER2 status should receive HER2 targeted therapies (Giordano, 2014). Results of HER2 testing are imperative to receive guideline concordant treatment. Studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010).

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Lund, M. J., E. N. Butler, et al. (2010). "Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes: a population-based study and first report." Cancer 116(11): 2549-2559.

Numerator Statement: HER2 testing performed
Denominator Statement: Adult women with breast cancer
Denominator Exclusions: None

Measure Type: Process
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Oct 22, 2012 **Most Recent Endorsement Date:** Oct 22, 2012

Maintenance of Endorsement -- Preliminary Analysis

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 *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.

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

- The developer provided evidence that addressed the correlation between HER2 status and benefit from anti-HER2 therapy.
- The developer provided a summary of the recommendations from the American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer:
    - The Panel recommends that HER2 status should be determined for all invasive breast cancer. A testing algorithm that relies on accurate, reproducible assay performance, including newly available types of bright-field ISH, is proposed. Elements to reliably reduce assay variation (for example, specimen handling, assay exclusion, and reporting criteria) are specified. An algorithm defining positive, equivocal, and negative values for both HER2 protein expression and gene amplification is recommended: a positive HER2 result is IHC staining of 3 (uniform, intense membrane staining of 30% of invasive tumor cells), a fluorescent in situ hybridization (FISH) result of more than six HER2 gene copies per nucleus or a FISH ratio (HER2 gene signals to chromosome 17 signals) of more than 2.2; a negative result is an IHC staining of 0 or 1, a FISH result of less than 4.0 HER2 gene copies per nucleus, or FISH ratio of less than 1.8. Equivocal results require additional action for final determination. **Level of Evidence:** Recommendation has not been graded.
- In 2012, the Steering Committee noted that HER2 testing is both prognostic and predictive of patient response to treatment therapies.

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 2013 guideline cited updated the one published in 2007.

Exception to evidence – In the absence of empirical evidence to support this process measure, the Committee may consider an exception to the evidence requirement with adequate justification.

Guidance from the Evidence Algorithm: Process measure(Box 1) → Systematic review conducted (Box 3)→ QQC provided (Box 4) --> SR concludes thatall women with breast cancer be tested for HER2 → high

Questions for the Committee:

- *For possible exception to the evidence criterion:*
    - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
    - *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*

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.

The developer provides the following information performance rates from the QOPI® registry:

	2013	2014	2015
# of practices	230	225	266
Total # of patients	6258	5980	6783
Overall	96.94	96.93	98.86
Mean	98.53	98.77	98.63
Min, Max	68.75, 100	80, 100	50, 100
Standard Deviation	3.31	2.72	4.26
Percentiles			
P10	95	96	96.2
P50	100	100	100
P90	100	100	100
P95	100	100	100

- In 2012, the Steering Committee members expressed concern with the presented performance gap stating concordance of 98 percent with the measure and questioned the opportunity for improvement. The developer noted that the participants on the measure are a self-selected group participating in the Quality Oncology Practice Initiative and performance

Disparities:

- The developer provided the following data on disparities:

	2013	2014	2015
Total # of patients	6,258	5,980	6,783
Overall	96.94	96.93	98.86
Hispanic	97.69	97.08	98.13
White	96.92	97.10	98.93
Black	96.08	94.96	98.92
Other	95.79	97.77	98.60

- The developer also noted that studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010).

Questions for the Committee:

- *Does a gap in care still exist that warrants a national performance measure?*
- *The developer presents some disparities data, are you aware of additional 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 to Support Measure Focus

Comments:

Process measure/clinician/group. Percent of women with breast cancer who are tested fro HER2 overexpression or amplification. Systematic review of the literature demonstrates the importance of HER2 status for appropriate therapeutic counseling. The evidence has not been graded.
The evidence reflects a specific process measure necessary for the appropriate classification and treatment of women with breast cancer. It is a direct measure, and its measurement impacts the treatment and prognosis of affected women. The desired outcome is improved survival. Measurement of HER 2 assists in the selection of appropriate treatment, which in turn has been shown to improve survival. If measuring a health outcome or PRO: is the relationship between the measured outcome/PRO and at least one healthcare action (structure, process, intervention, or service) identified AND supported by the stated rationale? Although the measure does not directly measure a health outcome (survival), it is a proxy for same. It does dictate choice of appropriate treatment (intervention/service) and is supported in the stated rationale. In response to the committee questions: The other measurement option would be overall survival, however this is neither practical nor in many instances achievable. Measuring treatment selection theoretically could be an alternate measure, but that choice must be informed by the results of the HER2 test. There could well be circumstances where the treatment choice was not informed by the test, and therefore not a reasonable quality indicator. The developers note that other guideline setting organizations such as NCCN have made recommendations regarding this test as well. There is a substantial peer reviewed literature on the value of HER2 testing as a prognostic and treatment indicator. There is RCT evidence of the value of HER2 testing. It is certainly acceptable and beneficial to hold providers accountable for this process of care. An exception is appropriate if necessary.

1b. Performance Gap
Comments:
QOPI registry reveiwed for 2013, 2014 and 2015 deomstrating consistently greater than 98% compliance with the measure. Range of compliance for 2015 was 50-100%--the largest in recent years. Although there was a suggestion of decreased complaince for AA in 2014, no differences were observed in 2015.
The data provided by the developer (QOPI) reflects a high performance rate for the measure. There is also evidence presented regarding disparities, and again performance is high across all groups. However the developer has also provided peer reviewed literature indicating nationwide performance gaps based on ethnicity. Therefore it is assumed—and probably correct—that the practices surveyed regarding performance were in a sense "self-selected" for improved performance given their willingness to participate in a practice improvement program. Committee questions: The developer presents evidence that a significant disparity related gap persists. Therefore a national performance measure is warranted. The American Cancer Society publishes data on disparities in cancer care, including breast cancer. Significant differences in outcomes in breast cancer survival based on ethnicity and socioeconomic status persist in the United States. Although all differences in survival in breast cancer are certainly not related to measurement of HER2, it remains an important part of quality cancer care. It should be measured.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic clinical data: registry

Specifications:

- This is a clinician-level measure.
- The numerator of this measure is: HER2 testing performed
- The denominator of this measure is: adult women with breast cancer
- ICD-10 codes are included; ICD-9 to ICD-10 conversion available on SharePoint.
- The calculation algorithm is provided.
- Instructions for obtaining a minimum sample size are provided.
- The developer specifies how missing data are handled.

- The developer stated that title and description were modified to clarify the measure intent and no other substantive changes were made. However, the following exclusions were in the previously endorsed version of this measure:
    - Patient history of metastatic cancer
    - Multiple primaries prior to or within the measurement period
- In 2012, Steering Committee members questioned whether patients with small tumor sizes should be excluded from the measure. The developer noted that insufficient sample size, as would result from a small tumor size, is included as a data element within the numerator. Further, the workgroup members agreed that an explicit exclusion of small tumor sizes may wrongly imply that HER2 testing on them is not necessary.

Questions for the Committee :

- *Are the appropriate codes included in the ICD-9 to ICD-10 conversion? Are all appropriate codes included?*
- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- The dataset used included 264 patient records from 44 QOPI practices submitted in spring 2007. Trained, independent nurse abstractors from the Virginia Quality Health Center served as the 'gold standard' against which practice abstractions were compared for accuracy.

Describe any updates to testing

- The developer indicated no updates to testing.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Current NQF reliability testing requirements include statistical analysis of the computed measure score or the individual patient-level data for the measured entities to determine the proportion of variation due to true differences vs. noise or random variation. Comparing practice abstractions against a 'gold standard' is considered data element validity testing.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing not conducted (Box 2)→Empirical validity testing of patient-level data conducted (Box 3) → Validity testing conducted with patient-level data elements (Box 10) →Appropriate method used but kappa scores for all data elements not provided →Insufficient

Questions for the Committee:

- *See questions under Validity*

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

Rationale: Insufficient information provided about the data elements tested and results.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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, summarize the validity testing from the prior review:

- Face validity of the measure score was assessed with input from experts involved in ASCO committees in 2011. The developer noted that the face validity survey results revealed that 95% of respondents strongly agreed or agree that this measure provided an accurate reflection of quality and can be used to distinguish good and poor quality. The developer did not provide the number of experts surveyed.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- Validity testing was conducted at the patient-level data element for 264 patient records using trained, independent nurse abstractors as the 'gold standard'. Kappa statistics were used to analyze the validity of the audited patient records compared to the submitted patient records. Kappa is the measure of agreement between two raters that adjusts for chance agreements for categorical data. Kappa values range between 0 and 1 and are interpreted as degree of agreement beyond chance. By convention, a kappa > .70 is considered acceptable.

Validity testing results:

- The developer provided a kappa score of 0.85. While this kappa score is above what is considered acceptable, the developer did not state which of the data elements this kappa score represents; no additional results were provided. NQF guidance states that testing should be done for all critical data elements.
- The developer did not state how it was determined which sampled patients met the denominator inclusion criteria. It is likely these were checked to ensure inclusion in the registry, but no results were reported.

Questions for the Committee:

- *Does the measure adequately identify and include HER2 positive breast cancer patients in the registry?*
- *Is the testing information provided enough to demonstrate sufficient validity so that conclusions about quality can be made?*
- *No updated testing information was presented. Does the Committee think there is a need to re-vote on validity?Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

N/A

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance*

measure scores can be identified):

The developer reported:

- Data from thefall 2011 QOPI round, data was submitted October and November 2011. 208 practices reported this measure. Data from 7987 patient records were submitted for this measure.

Year	Mean	St. Dev.	Min	10th	25th	50th	75th	90th	Max
2013	98.53	3.31	68.75	95	97.56	100	100	100	100
2014	98.77	2.72	80	96	97.74	100	100	100	100
2015	98.63	4.26	50	96.2	100	100	100	100	100

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

The developer indicates N/A.

2b7. Missing Data

The developer provides the following information:

- This measure is specified with defined criteria and data elements. If a patient record does not include one or more of these components for the initial patient population or denominator, then patients are not considered eligible for the measure and not included.
- If data to determine whether a patient should be considered for the numerator or exclusions is missing, then the numerator or exclusions not considered to be met and the practice will not get credit for meeting performance for that patient.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1) → potential threats to validity mostly addressed (Box2) → Empirical validity testing conducted (Box 3) →Validity testing not conducted at the measure score (Box 6)→Validity testing conducted with patient-level data elements (Box 10) →Appropriate method used but kappa scores for all data elements not provided →Insufficient

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**
Rationale: Insufficient information provided about the data elements tested and results.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Data elements clearly defined and appropriate codes included.

Committee questions: I do not have access to ICD9 and 10 codebooks so cannot respond to the question. The developers do provide code sets for inclusion. The data elements are clearly defined and reflect a complexity of determining a response to the measure. The logic algorithm is clearly stated. The complexity of determining the status of HER2 +/- based on the definitions provided reflects an inherent complexity of responding to the measure, along with the timelines as well as office locations. This will require chart review to complete for every patient, adding to administrative burden. This is not an e-measure, and the data assessments are possibly beyond the current capabilities of most systems to respond. Essentially, although at first glance it would appear a simple determination of numerator/denominator, it will in fact require a modest effort to report this measure. I have a concern that in practices that are not sophisticated in quality reporting, accuracy may suffer. Having said that, it is a measure we must continue to monitor.

2a2. Reliability Testing
Comments:
Data element testing initially. no further information provided.

2b2. Validity Testing
Comments:
Face validity reportedly 0.85, though no further data provided.
I have a great deal of respect for what QOPI has done, and the lengths to which they have gone to measure quality oncology care, and more specifically validate this measure. I am a bit more sanguine as to how this will be measured/reported in practices that may not have as strong a focus on quality. Determining validity in a more random practice sample to determine concordance would offer a higher degree of comfort that the "time" and "location" criteria specified in the measure would have been accurately reported. This will obviously be a matter discussed further by the committee.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Missing data for denomitor: patient not included. Missing data for numberator: measure failure. No exclusions are outlined.
2b3: N/A. 2b5: Although the data provided do not reflect significant differences in quality, peer reviewed literature does show significant differences across the country. It is unlikely that all practices meet the high level of compliance reflected by those who participate in the QOPI program. 2b6: No information available. 2b7: No, insofar as the developer states that such missing data will result in the practice not getting credit for meeting performance in that patient.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- All data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS). If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
- Data elements are generated or collected by and used by healthcare personnel during the provision of care.
- In 2012, the Steering Committee raised concerns that extraction of this data may be burdensome as it may require chart abstractions. Eventual use of this measure through EHRs would lessen this burden.

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Data points easily obtained from existing database/claims data.
**My question is whether the data is present in EHRs in a simple "yes/no" format, or does it require the more granular reporting described by the developers when determining whether the measure has been met? Also, there is a complex set of instructions about how often the patient has visited the practice, the time frame for reporting, the location of

offices, etc. Are there logic algorithms available in oncology EHRs which can assess the responses to these questions? I share the concerns about administrative burden, and whether the tools are in place to effectively respond to the measure criteria as outlined. If this was a simple "yes/no" from a date of first visit or a date of diagnosis, I would be more certain it could be accomplished. However, I need to better understand the burden of collecting this information. Parenthetically, this is a concern I have about quality measurement in general. As our systems become more sophisticated we should be able to reduce the reporting burden while improving the reporting and the quality. But are we there yet? So this is not an issue unique to this particular measure. It is more a reflection of our current capabilities in the typical oncology and medical practices nationwide.**

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and 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.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details
As reported by the developer, the measure is currently used in:

- Quality Oncology Practice Initiative:
- QOPI® Certification Program
- PQRS Qualified Clinical Data Registry

Improvement results

- The developer notes performance continues to remain high, some variation remains from year to year.
- Currently, performance results are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program. Data from reporting in the CMS PQRS program is not yet available but may provide additional information on overall performance scores on this measure in the near future.

Unexpected findings (positive or negative) during implementation: The developer reports no challenges or unexpected findings in implementation.

Potential harms: The developer reports no unintended consequences were noted during testing.

Feedback :

- MAP has not reviewed this measure for inclusion in any federal program.
- During the endorsement process in 2012, the Steering Committee, NQF members, and the public provided the following comments:
    - The Steering Committee members expressed concern that several measures had high rates of performance, indicating a small gap in performance; however, the developer clarified that the performance gap data came from the American Society for Clinical Oncology's Quality Oncology Practice Initiative (QOPI), which included self-selecting practices voluntarily reporting on measures. As such, the developer stated that it is likely that there is more variation in performance than was demonstrated through QOPI. The Steering Committee agreed with the developer that it is likely that there is variation in use of trastuzumab and in HER2 testing, given the self-selecting nature of the practices participating with QOPI. Taken in conjunction with several studies suggesting overuse of trastuzumab, the Steering Committee recommended the measure for endorsement.

- Public and member comments included:
  - A recommendation that a HER2 composite measure be developed, comprised of measures 1857, 1855, 1858, and 1878.
  - A recommendation that exclusion of de novo patients from testing to determine HER2 status be removed. The developer stated that the measure does not recommend against testing among patients who are excluded from the denominator (patients with metastatic disease or multiple primaries prior to or within the measurement period). Future development work could consider measurement to address HER2 re-testing, if supported sufficiently by evidence and if feasibility/burden were considered appropriate.
  - The developer stated that ASCO and CAP, the developers of the referenced measures, have discussed the concept of a composite measure, and neither organization believes that it is advantageous at this time. These measures are designed for different providers and levels of accountability, and have different denominators. Measure 1855 was developed to measure the performance of individual pathologists, while measures 1857, 1858, and 1878 are for medical oncologists/clinical oncology practices. It may be beneficial to implement all of these measures within certain settings, such as accountable care organizations or Cancer Care Centers. ASCO reports measures 1857, 1858, and 1878 together in their quality programs; however, they believe that the measures are independently useful. The developer will consider paired or composite measures in the future. The Steering Committee agreed that as the measures are currently specified for different levels of analysis, a composite measure would not be feasible. Further, the Steering Committee agreed that the measures capture discrete steps in care.

***Questions for the Committee**:*

- *When publicly reported, how useful is it to patients in making comparisons?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Although the measure is in use, is the small (to no) improvement over time indicative of poor usability?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Currently used for accountability in QOPI and PQRS. Performance is very high.

**Committee questions: This measure is very useful when publicly reported. It is essentially a universally accepted standard of care, and consumers should be aware of any practice which provides a less than satisfactory compliance with this measure. This test is a basic element of breast cancer care, and dictates the treatment needed to effectively care for women diagnosed with breast cancer. As such, it is part of the basic criteria of a high quality cancer care system.
I have not seen more recent data relative to disparities in care delivery, so I cannot comment whether the performance has improved since this measure is available. I do not believe that the high compliance reflected in the data supplied by the measure's developer is any indication of poor usability. Having said that, I did comment earlier about concerns regarding reporting variation based on some of the complexities of data interpretation. I doubt that we have information from oncology practices on the administrative burden of reporting this measure, but perhaps that is a question that should be asked by the users/reporters.**

Criterion 5: Related and Competing Measures

Related or competing measures

- 1855 : Quantitative HER2 evaluation by IHC uses the system recommended by the ASCO/CAP guidelines
  - 1855 and 1878 address two complimentary components related to appropriate identification and treatment of breast cancer patients.
  - 1855 and 1878 differ by data source. Measure #1878 is suited for registry data. Measure #1855 is suited

for administrative claims and paper medical records data sources.

Harmonization

- The developer indicates the measures have been harmonized

Pre-meeting public and member comments

-

NQF #: 1878 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
This is a process measure that is closely linked to outcomes for invasive breast cancer patients. HER2 overexpression is associated with clinical outcomes in patients with breast cancer. HER2 positivity is associated with worse prognosis (higher rate of recurrence and mortality) in patients with newly diagnosed breast cancer who do not receive any adjuvant systemic therapy. Thus, HER2 status might be incorporated into a clinical decision, along with other prognostic factors, regarding whether to give any adjuvant systemic therapy. HER2 status is also predictive for several systemic therapies, including trastuzumab.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The primary outcome of interest was the correlation between HER2 status and benefit from anti-HER2 therapy. Other outcomes of interest included the positive predictive value (PPV) and negative predictive value (NPV) of fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC) to determine HER2 status, alone and in combination; concordance across platforms; and accuracy in determining HER2 status, sensitivity, and specificity of specific tests.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: The following electronic databases were searched from January 1987 through February 2006: MEDLINE, PreMEDLINE, and the Cochrane Collaboration Library. In addition, abstracts presented at ASCO or CAP from 2000 to 2005 and at the San Antonio Breast Cancer Symposium from 2003 to 2005 were identified. Results were supplemented with hand searching of selected reviews and personal files.

Preliminary searches identified 1,802 MEDLINE abstracts. The initial abstract screen performed by ASCO staff eliminated 1,010 abstracts that failed to meet any of the inclusion criteria. The ASCO panel conducted dual independent review of all remaining 792 potentially relevant abstracts identified in the original systematic review. The panel eliminated 667 abstracts at this stage of the review; the remaining 125 articles were reviewed in full for the interventions and outcomes described herein.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* Articles were selected for inclusion in the systematic review of the evidence if they met the following criteria: (1) the study compared, prospectively or retrospectively, the negative predictive value (NPV) or positive predictive value (PPV) of FISH or IHC; the study described technical comparisons across various assay platforms; the study examined potential testing algorithms for HER2 testing; or the study examined the correlation of HER2 status in primary versus metastatic tumors from the same patients; and (2) the study population consisted of patients with a diagnosis of invasive breast cancer; and (3) the primary outcomes included the PPV and NPV of FISH and IHC to determine HER2 status, alone and in combination; concordance across platforms; accuracy in determining HER2 status and benefit from anti-HER2 therapy, sensitivity, and specificity of

specific tests. Consideration was given to studies that directly compared results across assay platforms. Evidence tables were developed based on selected studies that met the criteria for inclusion.

Study design was not limited to randomized controlled trials, but was expanded to include any study type, including cohort designs, case series, evaluation studies, comparative studies, and prospective studies. Also included were testing guidelines and proficiency strategies of various countries, primarily the United States, and international organizations. Letters, commentaries, and editorials were reviewed for any new information. Case reports were excluded.

1c.7 **Consistency of Results <u>across Studies</u>** *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**
Based on preliminary reports of three large RCTs, the addition of one year of trastuzumab, following a variety of adjuvant or neoadjuvant chemotherapy regimens, significantly improved the primary endpoint of DFS in patients with HER2/neu positive early breast cancer. Secondary endpoints of RFS, DDFS, and TTR in all studies, and OS in one combined study, were also significantly improved with the addition of trastuzumab. Those results are only applicable to women with HER2/neu overexpressing breast cancer who complete a minimum of four cycles of adjuvant or neoadjuvant chemotherapy. Although the majority of the patients in those studies had node-positive breast cancer, women with high-risk node-negative breast cancer were also included in HERA (32% were N0 but had tumours T1c) and NCCTG 9831 (11% were N0 but had tumours >1cm if ER negative, >2cm if ER positive). Therefore, those results are also generalizable to women with node-negative breast cancer meeting these criteria. The magnitude of incremental benefit conveyed by adjuvant trastuzumab well exceeds the gains accrued by over three decades of adjuvant chemotherapy use.

Based on experience in the metastatic setting, the concurrent use of trastuzumab and anthracyclines has prohibitive cardiac toxicity. Based on the current reports, the cardiac toxicity with adjuvant trastuzumab appears to be acceptable, although the reported rate of cardiac events was higher in the concurrent versus sequential trastuzumab arm (in NSABP B31 4.1% vs. 0.7%, HR of 7.2; in NCCTG 9831 3.3% vs. 2.2%). The toxicity is considered acceptable, given the increase in survival.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** The guideline panel provided a narrative review of the body of evidence.

For the 2012 Update, in process (published in 2013) the new "ASCO Summary Ratings including the Assessment of Study Quality, Strength of Evidence, and Strength of Recommendations System" is being used.

The ASCO Health Services Committee (HSC) and the CAP Council on Scientific Affairs (CSA) jointly convened an expert panel consisting of experts in clinical medicine and research relevant to HER2 testing, including medical oncology, pathology, epidemiology, statistics, and health services research. Academic and community practitioners and a patient representative were also part of the panel. Representatives from the US Food and Drug Administration, the Centers for Medicare and Medicaid Services, the National Cancer Institute, and the National Academy of Clinical Biochemistry served as ex-officio members .The opinions of panel members associated with official government agencies represent their individual views and not necessarily those of the agency with which they are affiliated.

All members of the expert panel complied with ASCO policy on conflict of interest, which requires disclosure of any financial or other interest that might be construed as constituting an actual, potential, or apparent conflict. Members of the expert panel completed ASCO's disclosure form and were asked to identify ties to companies developing products that might be affected by promulgation of the guideline. Information was requested regarding employment, consultancies, stock ownership, honoraria, research funding, expert testimony, and membership on company advisory committees. The panel made decisions on a case-by-case basis as to whether an individual's role should be limited as a result of a conflict. No limiting conflicts were identified.

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Detailed narrative description of the strength of each study.

1c.13 **Grade Assigned to the Body of Evidence:** n/a

1c.14 Summary of Controversy/Contradictory Evidence: A thorough discussion of the limitations of the literature and/or controversies is included in the "Summary and Recommendations" section in the body of the guideline, presented after each Clinical Question.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
Verbatim recommendations can be viewed at http://jco.ascopubs.org/content/25/1/118.full.pdf in Table 4 on page 123.

Summary of Recommendations:

The Panel recommends that HER2 status should be determined for all invasive breast cancer. A testing algorithm that relies on accurate, reproducible assay performance, including newly available types of bright-field ISH, is proposed. Elements to reliably reduce assay variation (for example, specimen handling, assay exclusion, and reporting criteria) are specified. An algorithm defining positive, equivocal, and negative values for both HER2 protein expression and gene amplification is recommended: a positive HER2 result is IHC staining of 3 (uniform, intense membrane staining of 30% of invasive tumor cells), a fluorescent in situ hybridization (FISH) result of more than six HER2 gene copies per nucleus or a FISH ratio (HER2 gene signals to chromosome 17 signals) of more than 2.2; a negative result is an IHC staining of 0 or 1, a FISH result of less than 4.0 HER2 gene copies per nucleus, or FISH ratio of less than 1.8. Equivocal results require additional action for final determination.

1c.17 Clinical Practice Guideline Citation: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer.
Wolff, A.C, Hammond, M.E.H, et.al., "Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update." J Clin Onc 31.31 (2013): 3997-4013. Available at: http://jco.ascopubs.org/content/31/31/3997.full

1c.18 National Guideline Clearinghouse or other URL: http://www.guideline.gov/content.aspx?id=10384&search=her2+testing AND http://jco.ascopubs.org/content/25/1/118.full.pdf+html

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? No

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: n/a

1c.23 Grade Assigned to the Recommendation: n/a

1c.24 Rationale for Using this Guideline Over Others: The systematic review and the process of rating the body and evidence and the strength of recommendations makes it a very transparent and credible document. The collaboration between ASCO and the College of American Pathologists (CAP) assured a consistent message to medical oncologists and pathologists. Other guidelines recommend HER2 testing for women with invasive breast cancer (e.g., National Comprehensive Cancer Network).

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1878_Evidence_MSF5.0_Data-635932990174672035.doc

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., the benefits or improvements in quality envisioned by use of this measure)*
Human epidermal growth factor receptor (HER2) gene is amplified and/or overexpressed in approximately 15% to 20% of primary breast cancers. The ASCO/CAP joint guideline on HER2 testing recommends all patients with invasive breast cancer should be tested for HER2 status and only those who test positive for HER2 status should receive HER2 targeted therapies (Giordano, 2014). Results of HER2 testing are imperative to receive guideline concordant treatment. Studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010).

Giordano, S.H., Temin, S., et. al., "Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline." J Clin Onc 32.19 (2014): 2078-099. Available at:http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Lund, M. J., E. N. Butler, et al. (2010). "Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes: a population-based study and first report." Cancer 116(11): 2549-2559.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This data was produced from the QOPI® registry and data was abstracted for a sample of patients seen with the data collection period. Performance is reported at the clinical practice level.

In 2013, 230 practices were measured, and the total patient population for this measure was 6258.

In 2014, 225 practices were measured, and the total patient population for this measure was 5980.

In 2015, 266 practices were measured, and the total patient population for this measure was 6783.

	2013	2014	2015
Overall	96.94	96.93	98.86
Mean	98.53	98.77	98.63
Minimum	68.75	80	50
Maximum	100	100	100
Standard Deviation	3.31	2.72	4.26
Percentiles			
P10	95	96	96.2
P25	97.56	97.74	100
P50	100	100	100
P75	100	100	100
P90	100	100	100
P95	100	100	100

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.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
This data was produced from the QOPI® registry and data was abstracted for a sample of patients seen with the data collection period. Performance is reported at the chart level.

In 2013, the total patient population for this measure was 6258.
In 2014, the total patient population for this measure was 5980.
In 2015, the total patient population for this measure was 6783.

	2013	2014	2015
Overall	96.94	96.93	98.86
Hispanic	97.69	97.08	98.13
White	96.92	97.10	98.93
Black	96.08	94.96	98.92
Other	95.79	97.77	98.60

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010).

Health disparities between patients with breast cancer according to race/ethnicity, age, insurance status, geographic location, education, and other factors are well documented, however literature addressing disparities specific to patients with HER2-positive metastatic breast cancer is scarce. According to some studies, there are not large (although some suggest modest) differences in the prevalence of HER2 positivity between women with breast cancer of different races/ethnicities. The variation by race is smaller among those with HER2-positive breast cancer than for some other subtypes.

HER2 positivity is not necessarily associated with worse treatment outcomes among African American compared with non–African American patients. However, high-quality data on patients with HER2-positive metastatic disease are still needed to reach conclusions related to outcomes based on ethnicity. Therefore, health disparities may be similar to those faced by patients with metastatic breast cancer generally.

Although ASCO clinical practice guidelines represent expert recommendations the highest level of cancer care, it is important to note that many patients have limited access to medical care. Racial and ethnic disparities in health care contribute significantly to this problem in the United States. Minority racial/ethnic patients with cancer suffer disproportionately from comorbidities, experience more substantial obstacles to receiving care, are more likely to be uninsured, and are at greater risk of receiving care of poor quality than other North Americans. Many other patients lack access to care because of their age, geography, and distance from appropriate treatment facilities. Awareness of these disparities in access to care should be considered in the context of this clinical practice guideline, and health care providers should strive to deliver the highest level of cancer care to these vulnerable populations (Giordano, 2014).

Citations:

Giordano, S.H., Temin, S., et. al., “Systemic Therapy for Patients with Advanced Human Epidermal Growth Factor Receptor 2- Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline.” J Clin Onc 32.19 (2014): 2078-099. Available at: http://jco.ascopubs.org/content/32/19/2078.full.pdf+html

Lund, M. J., E. N. Butler, et al. (2010). "Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes: a

population-based study and first report." Cancer 116(11): 2549-2559.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*
Cancer, Cancer : Breast

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

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)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Title and description were modified to clarify the measure intent.

No other substantive changes were made.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
HER2 testing performed

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Practices are required to order tests within 31 days from first office visit (HER2 test date – first office visit date = 31 days) and if a new test is ordered, it must be within 10 days of original report

S.6. **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, 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.
HER-2/neu status = HER2 positive
OR
HER-2/neu status = HER2 negative
OR
HER-2/neu status = Test ordered, results not yet documented
OR

HER-2/neu status = Test ordered, insufficient sample for results
OR
(HER-2 equivocal AND New test ordered within 10 days of report = Yes or N/A (patient died or transferred out of practice))

Practices are required to order tests within 31 days from first office visit (HER2 test date – first office visit date = 31 days) and if a new test is ordered, it must be within 10 days of original report

Numerator definitions:
Select 'Test ordered, results not yet documented' only if there is documentation in the chart that a test that reports HER-2/neu analyses was ordered.

In the absence of any documentation regarding HER-2/neu status, select 'Test not ordered/no documentation.'
Enter information from the most recent test report.
Patients are classified as having HER-2 positive disease based on positive results with either test.
If the most recent report indicates insufficient sample, select 'Test ordered, insufficient sample for results.'
If a physician note and the HER-2/neu report differ in results, report the status in the physician note if the note explains the discrepancy. Otherwise, report the status from the HER-2/neu report.

Use the following definitions to determine HER-2/neu status:
Positive:
- IHC 3+ cell surface protein expression (defined as uniform intense membrane staining of >30% of invasive tumor cells) or
- FISH ratio >2.2 or
- HER2 gene copy >6.0

Equivocal:
- Not positive according to any of the criteria above, AND
- (IHC with scores 2+ AND FISH ratio 1.8-2.2) or
- HER2 gene copy 4.0-6.0

Negative:
- Not positive according to any of the criteria above, AND
- IHC 0 or 1+ or
- FISH ratio 1.8 or
- HER2 gene copy <4.0

- If the results indicate 'non-amplified', choose HER-2/neu negative.
- If the results indicate 'weakly positive', choose HER-2/neu positive.

New test ordered within 10 days of report of equivocal result: Respond 'Yes' if a new test was ordered within 10 days of oncologist review of the report with inconclusive results. Choose 'N/A' if the patient died or transferred out of the practice within 10 days of review of the report with inconclusive results or fewer than 10 days have passed.
If the chart documents that the pathologist has ordered a new test, respond 'Yes.'

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Adult women with breast cancer

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Female
And
2 or more encounters at the reporting site
And

Age at diagnosis greater than or equal to 18 years
And
Breast cancer diagnosis [C50.01-, C50.11-, C50.21-, C50.31-, C50.41-, C50.51-, C50.61-, C50.81-, C50.91-]

Definitions

Encounter: Patients must have been first seen in the office by a medical oncology or hematology oncology practitioner for the cancer diagnosis eligible for inclusion within the 1-year time frame of the reporting period. Enter the most recent visit that occurred during the 6-month visit window before the abstraction date. This can include visits to other office sites within the practice only if the practice uses a common medical record and shares management of care for the patient. This does not include visits during which a practitioner wasn´t seen (e.g., laboratory testing), inpatient consults/visits, phone or email consults, or visits to a surgeon or radiation oncologist.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
None

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not applicable

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
Performance is calculated as:
1. Identify those patients that meet the denominator criteria defined in the measure.
2. Subtract those patients with a denominator exclusion from the denominator. Note: this measure does not have exclusions.

3. From the patients who qualify for the denominator (after any exclusions are removed), identify those who meet the numerator criteria.
4. Calculation: Numerator/Denominator-Denominator Exclusions

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
QOPI registry abstraction is offered twice a year to participating Medical Oncology and Hematology Oncology Practices. The minimum sample size for each data abstraction period is based on the number of med-onc and hem-onc FTEs at the practice and/or site level. For breast cancer, patients must be female, 18 years and older with a diagnosis of [C50.01-, C50.11-, C50.21-, C50.31-, C50.41-, C50.51-, C50.61-, C50.81-, C50.91-] within the one year diagnosis window applicable to the round. The practices follow a chart selection methodology which identify patients who had a diagnosis date within one year of the abstraction period start date AND had two office visits with a practitioner in the office within three months of the abstraction data period start date.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
This measure is specified with defined criteria and data elements. If a patient record does not include one or more of these components for the initial patient population or denominator, then patients are not considered eligible for the measure and not included.
If data to determine whether a patient should be considered for the numerator or exclusions is missing, then the numerator or exclusions not considered to be met and the practice will not get credit for meeting performance for that patient.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
ASCO Quality Oncology Practice Initiative (QOPI®)

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1878_MeasureTesting_MSF5.0_Data_Update.doc

NQF #: 1878 **NQF Project:** Cancer Project

2. RELIABILITY & 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. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
2008 IRR study: A sample of 300 records was planned for re-abstraction in four geographic regions: Midwest, Northeast, South, and West. 50 QOPI practices were randomly selected from the 4 geographic areas and invited to participate. Within each practice, six previously abstracted charts were selected randomly for re-abstraction from the population of 13,561 records submitted in spring 2007 round. Forty-four practices agreed to participate, and submitted 264 records (6 per practice).

2010-2011 audit: QOPI practices applying for the QOPI Certification Program are required to submit copies of documentation from 3-5 records which were previously abstracted. Trained ASCO auditors randomly select records within each relevant module for audit. Agreement at the data element level is documented. 426 audited records from 130 practices were complete in November 2011 and included in the concordance analysis.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
2008 IRR study: ASCO engaged the Virginia Quality Health Center to conduct an inter-rater reliability study of the QOPI case report form and measures. Trained, independent nurse abstractors served as the 'gold standard' against which practice abstractions were compared for accuracy. Sampling is described above. The 264 sampled records allowed for reliability analysis at a 95% confidence level with a +/- 3.88% marking of error.

Kappa statistics were used to analyze the reliability of the audit data set compared to the submitted data. Kappa statistics are the commonly accepted standard for determining inter-rater reliability in the healthcare setting (Allison, Calhoun, et al, 2000; Cassidy, Marsh, et al, 2002). The Kappa statistic is conceptually similar to the rate of agreement between two reviewers, but it imposes a more stringent standard that simple agreement and mismatch rates. The following standards were used (Cohen, 1960; Sim and Wright, 2005; Feinstein and Cicchetti, 1990):

- Kappa > .0.75 denotes excellent reliability
- Kappa between 0.40 and 0.75 denotes good reliability
- Kappa less than 0.40 denote marginal reliability

2010-2011 audit: Agreement data from 426 records were imported into a formatted data table for analysis. First, agreement data were used to calculate concordance at the data element level. Second, by applying the measure analytic calculation, concordance at the measure level was calculated.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
2008 IRR study: measure level Kappa 0.85 (excellent reliability)

2010-2011 audit: measure level concordance 98% (valid N=132 records)

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
Measure specifications are consistent with the evidence cited. Members of the ASCO guideline development panel reviewed the measure specification to ensure consistency.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b2.2 **Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
In 2009, an ASCO steering group comprised of medical oncologists, health services researchers, and quality experts undertook an iterative, criteria-based assessment process to identify QOPI measures that are appropriate for use for accountability measurement. This measure was selected as appropriate for accountability.

Face validity of the measure score was assessed via survey of experts involved in ASCO committees in 2011. The survey explicitly asked whether the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
Face validity survey results revealed that 95% of respondents 'strongly agree' or 'agree' that this measure provides an accurate reflection of quality and can be used to distinguish good and poor quality.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
n/a

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
n/a

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
n/a

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
n/a

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
n/a

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
n/a

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**

n/a

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Data reported are from the Fall 2011 QOPI round, reflecting data submitted October and November 2011. 208 practices reported this measure. Data from 7987 patient records were submitted for this measure.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
QOPI measure analytics at the practice level were generated. Practices with fewer than 5 records were not included in calculations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

Year	Mean	St. Dev.	Min	10th	25th	50th	75th	90th	Max
2013	98.53	3.31	68.75	95	97.56	100	100	100	100
2014	98.77	2.72	80	96	97.74	100	100	100	100
2015	98.63	4.26	50	96.2	100	100	100	100	100

This measure has been implemented in QOPI for several years. In this self-selected group on oncology practitioners committed to quality assessment and improvement, concordance with this measure has been high overall; however, a proportion of practices (new and experienced) continue to demonstrate sub-optimal variation.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
n/a

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**
n/a

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**
n/a

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)***:** n/a

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
n/a

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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)*

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

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure and its specifications have been in place for several years and ASCO continues to monitor and ensure that the measure and its specifications are up-to-date for widespread use.

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

Not applicable

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Planned	Current Use (for current use provide URL)
	Payment Program CMS Physician Quality Reporting Program Qualified Clinical Data Registry http://www.instituteforquality.org/qopi/pqrs-measures-0 Professional Certification or Recognition Program QOPI® Certification Program http://www.instituteforquality.org/qcp/qopi-certification-measures Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Oncology Practice Initiative (QOPI®) http://www.instituteforquality.org/qopi/manual-qopi-measures

4a.1. For each CURRENT use, checked above, provide:

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

Quality Oncology Practice Initiative:
In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is a practice-based quality assessment and improvement program designed to foster a culture of self-examination and improvement in oncology. Collection rounds are offered twice per year, in spring and fall, for an eight week period. QOPI® continues to be a successful program in the United States and 12 other countries, with 441, 313, 361 and 256 unique practices participating in Fall 2013, Spring 2014, Spring 2015 and Fall 2015 respectively.

QOPI® Certification Program:
The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score. Please see a description of the QOPI® program above for details.

PQRS Qualified Clinical Data Registry:
In addition to the current use for quality improvement with benchmarking in the QOPI® registry, this measure has been reported to CMS by the registry as a Qualified Clinical Data Registry. QOPI® was deemed as a registry for oncology measures group reporting and as a QCDR to report to PQRS in 2015 and 2016. Eligible professionals will be considered to have satisfactorily participated in PQRS if they submit quality measures data or results to CMS via a qualified clinical data registry. In Fall 2015, 36 practices and 3,124 patient charts were submitted to PQRS through QOPI. 2015 QCDR data will be publicly reported at the individual eligible professional level as a performance rate in the form of a percent for each measure. Beginning with 2016 data, both individual and group-level QCDR performance rates will be publicly reported.

4a.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?*)
As described above, CMS is planning to publicly report QCDR data.

Additionally, although the measure is currently in use, we will continue to seek opportunities to advocate for expanded use of this measure in government or other programs.

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

While performance continues to be generally high, some variation still remains as evidenced by the performance ranges by year and from the time of the last NQF endorsement review. The data available are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program.

Studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010). In addition, data from reporting in the CMS PQRS program is not yet available but may provide additional information on overall performance scores on this measure in the near future.

Lund, M. J., E. N. Butler, et al. (2010). "Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes: a population-based study and first report." Cancer 116(11): 2549-2559.

4b.2. **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.**
While performance continues to be generally high, some variation still remains as evidenced by the performance ranges by year and from the time of the last NQF endorsement review. The data available are based on QOPI® self selecting practices that voluntarily report data and may not be reflective of care provided outside of the QOPI® program.

Studies show that tumors of older female patients (15.7%) and Hispanics (20.7%) as well as other race/ethnicities (18.8%) are less likely to be tested for HER2 (Lund, 2010). In addition, data from reporting in the CMS PQRS program is not yet available but may provide additional information on overall performance scores on this measure in the near future.

Lund, M. J., E. N. Butler, et al. (2010). "Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes: a population-based study and first report." Cancer 116(11): 2549-2559.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There have been no reports of unintended consequences with 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)**
1855 : Quantitative HER2 evaluation by IHC uses the system recommended by the ASCO/CAP guidelines

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

5a. Harmonization

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 completely harmonized?
Yes

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.)
Measure #1878 assesses whether HER2 testing was completed within 31 days of a breast cancer diagnosis. Meanwhile, NQF endorsed measure #1855 focuses on whether HER2 testing was completed according to current ASCO/CAP standards in the laboratory setting. As such, these measures address two complimentary components related to appropriate identification and treatment of breast cancer patients.

These measures also differ by data source. Measure #1878 is suited for registry data while Measure #1855 is suited for administrative claims and paper medical records data sources.

Because each measure has a different intent and uses a different data source, both measures should maintain their endorsement.

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:** QOPI_Adoption_of_ICD10_020916.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Society of Clinical Oncology
Co.2 **Point of Contact:** Tayyaba, Shehzadi, Tayyaba.Shehzadi@asco.org, 571-483-1673-
Co.3 **Measure Developer if different from Measure Steward:** American Society of Clinical Oncology
Co.4 **Point of Contact:** Tayyaba, Shehzadi, Tayyaba.Shehzadi@asco.org, 571-483-1673-

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.
ASCO Breast Cancer Measure Development Panel
The panel is responsible for reviewing and maintaining breast cancer measures

Gary Lyman, MD, MPH, FASCO, FRCP
Co-Chair
Fred Hutchinson Cancer Research Center

Gabrielle Rocque, MD
Co-Chair
University of Alabama

Banu Arun, MD
University of Texas
MD Anderson Cancer Center

Gary Cohen, MD, FASCO
Cancer Center at GBMC

Shelley Fuld Nasso, MPP
National Coalition for Cancer Survivorship

Jennifer Griggs, MD, MPH
University of Michigan

Michael Hassett, MD, MPH
Dana-Farber Cancer Institute

Michael Neuss, MD, FASCO
Vanderbilt Ingram Cancer Center

Ann Partridge, MD, FASCO
Dana-Farber Cancer Institute

Michael Soble, MD
North Shore Oncology/Hematology Associates

Ann Von Gehr, MD
Permanente Medical Group Inc.

Antonio Wolff, MD, FACP, FASCO
Johns Hopkins Kimmel Cancer Center

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 02, 2016
Ad.4 **What is your frequency for review/update of this measure?** q3years

Ad.5 **When is the next scheduled review/update for this measure?** 02, 2017

Ad.6 **Copyright statement:** The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.
The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.
Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] or American Society of Clinical Oncology (ASCO).
Neither the AMA, ASCO, PCPI, nor its members shall be responsible for any use of the Measures.
The AMA's and PCPI's significant past efforts and contributions to the development and updating of the Measures is acknowledged.
ASCO is solely responsible for the review and enhancement ("Maintenance") of the Measures as of 2016.
ASCO encourages use of the Measures by other health care professionals, where appropriate.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

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. The AMA, ASCO, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

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 #: 2930
Measure Title: Febrile Neutropenia Risk Assessment Prior to Chemotherapy
Measure Steward: RAND Corporation
Brief Description of Measure: Percentage of patients with a solid malignant tumor or lymphoma who had a febrile neutropenia (FN) risk assessment completed and documented in the medical record prior to the first cycle of intravenous chemotherapy
Developer Rationale: This process measure focuses on assessing the risk of febrile neutropenia (FN) in patients with a solid malignant tumor or lymphoma prior to receiving their first cycle of an intravenous chemotherapy regimen. FN is a complication of chemotherapy that occurs as a result of chemotherapy-induced neutropenia, causing the patient to be highly susceptible to infection. FN after chemotherapy occurs frequently, with the incidence of FN among patients with solid tumors estimated to be 13.1-20.6% during their chemotherapy course and 3.1-7.4% in the first cycle (Weycker et al., 2015), and incidence among patients with lymphoma estimated at 36% (Bohlius et al., 2008). If a patient presents with fever during the neutropenic phase, antibiotic treatment (usually intravenous) and often hospital admission are required to control a likely infection and prevent the development of sepsis, and other complications, including death. Estimates of mortality for patients who were hospitalized for complications related to FN range from 7 to 20 percent among those with solid tumors, with higher rates among those with comorbidities (Kuderer et al., 2006; Elting et al., 1997; Schwenkglenks et al., 2006; Segal et al., 2008), and 9 percent among those with lymphoma (Kuderer et al., 2006).

Having information about a patient's FN risk allows the identification of patients at higher risk of FN who are more likely to benefit from treatment with prophylactic colony-stimulating factor (CSF) which stimulates the production of white blood cells and lowers the risk of FN and its complications. If a higher proportion of patients are assessed for FN risk, more of those with a higher FN risk would receive CSF and a lower proportion of patients would be expected to develop FN and its complications.

Citations

Bohlius, J., Herbst, C., Reiser, M., Schwarzer, G., & Engert, A. (2008). Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. Cochrane Database Syst Rev(4), Cd003189.

Elting, L. S., Rubenstein, E. B., Rolston, K. V., & Bodey, G. P. (1997). Outcomes of bacteremia in patients with cancer and neutropenia: observations from two decades of epidemiological and clinical trials. Clin Infect Dis, 25(2), 247-259.

Kuderer, N. M., Dale, D. C., Crawford, J., Cosler, L. E., & Lyman, G. H. (2006). Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. Cancer, 106(10), 2258-2266.

Schwenkglenks M., J. C., Constenla M., Leonard R.C. (2006). Neutropenic event risk and impaired chemotherapy delivery in six European audits of breast cancer treatment. Supportive Care Cancer, 14(9), 901-909.

Segal, B. H., Freifeld, A. G., Baden, L. R., Brown, A. E., Casper, C., Dubberke, E., et al. (2008). Prevention and treatment of cancer-related infections. J Natl Compr Canc Netw, 6(2), 122-174.

Weycker, D., Li, X., Edelsberg, J., Barron, R., Kartashov, A., Xu, H., & Lyman, G. H. (2014). Risk and Consequences of Chemotherapy-Induced Febrile Neutropenia in Patients With Metastatic Solid Tumors. Journal of Oncology Practice, 11(1), 47-54.

Numerator Statement: Number of patients who had an FN risk assessment documented in the medical record prior to the first cycle of intravenous chemotherapy.
Denominator Statement: Number of patients 18 years of age or older with a solid malignant tumor or lymphoma receiving the first cycle of intravenous chemotherapy.
Denominator Exclusions: There are no denominator exclusions

Measure Type: Process
Data Source: Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

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 provided a diagram illustrating the path between assessing the risk of febrile neutropenia (FN) in patients with a solid malignant tumor or lymphoma prior to receiving their first cycle of an intravenous chemotherapy regimen and the proportion of patients receiving prophylactic colony-stimulating factor (CSF) and the rate of adverse patient complications.
- The developer provided two clinical practice guidelines that support risk assessment for chemotherapy-induced febrile neutropenia (FN):
  - 2015 American Society of Clinical Oncology (ASCO) Recommendations for the Use of WBC Growth Factors: Primary prophylaxis with a CSF starting with the first cycle and continuing through subsequent cycles of chemotherapy is recommended in patients who have an approximately 20% or higher risk for febrile neutropenia based on patient-, disease- and treatment-related factors. Primary CSF prophylaxis should also be administered in patients receiving dose dense chemotherapy when considered appropriate. Consideration should be given to alternative, equally effective, and safe chemotherapy regimens not requiring CSF support when available." (Emphasis added.) **Level of Evidence: Strong**
  - 2015 NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®): The guidelines begin with ***an evaluation of risk for chemotherapy-induced FN prior to the first cycle of chemotherapy***. The risk assessment includes disease type, chemotherapeutic regimen (high-dose, dose-dense, or standard-dose therapy), patient risk factors, and treatment intent. Three categories based on the intent of chemotherapy have been designated by the NCCN Panel. These include curative-adjuvant therapy, treatment directed toward prolongation of survival, and symptom management therapy. Based on the chemotherapy regimen and patient-related risk factors, the patient is assigned to either an overall high-risk group (>20% risk of FN), intermediate risk group (10%-20% risk), or low-risk group (<10% risk). Of note, there is currently no consensus nomogram for risk assessment. While the NCCN Panel outlines

criteria to aid in the assessment of FN risk, independent clinical judgment should be exercised based on the patient's situation (see Patient Risk Factors for Developing Febrile Neutropenia in the algorithm). In addition to assessing patient- and treatment-related risk, consideration should be given to the intent of cancer treatment when determining the appropriate use of CSFs. For example, a patient with a previous neutropenic complication in the immediately prior cycle of chemotherapy, with no plan to reduce the dose intensity should be considered high risk." (Emphasis added.) **Level of Evidence: Category 2A.** *[Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.]*

  - The developer also provided an algorithm of the 2015 Clinical Practice Guidelines on Myeloid Growth Factors (**NCCN Guidelines®**) illustrating the factorsing leading to the decision about prophylactic use of CSF for febrile neutropenia. **Level of Evidence: Category 2A.** *[Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.]*
- The developer provided the systematic review including a summary of the Quantity, Quality, and Consistency (QQC) of the body of evidence.
- The developer provided an additional seven articles as a source of evidence for this measure.

Exception to evidence: N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → SR with QQC (Box 4) → Quantity: High; Quality: High; Consistency: High (Box 5a) → High

Questions for the Committee:

- *Is the evidence directly applicable to the process of care being measured?*

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

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

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

The developer provided the following statistics for the measure rate by clinics:

No. of clinics	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
5	0.127	0.162	0	0.270	0.113	0.154	0.01	0.025	0.162	0.179	0.234

- The developer did not provide the dates of data and number of patients included in the statistics above, though list a total of 192 patients in the data presented for disparities below.
- The developer provided additional data from the literature.

Disparities:
The measure was stratified for disparities by age, race/ethnicity, and gender for the entire sample. The developer provided the following results/scores in the tables listed below:

	Denominator	Numerator	Measure Rate
All Patients	192	24	0.125
18 – 44	27	6	0.222
45 – 64	69	5	0.072
65 – 74	63	8	0.127
75 – 84	30	5	0.167
85+	3	0	0
White, non-Hispanic:	111	17	0.153
Black, non-Hispanic	16	0	0
Hispanic	30	2	0.067

Other	13	2	0.154
Unknown	22	3	0.136
Female	134	20	0.149
Male	58	4	0.069

- The developer provided additional disparities data from a summary from the literature.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Do you agree 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 to Support Measure Focus

Comments:

Excellent

**Evidence is there to support this measure and is applicable to the process of care. Two clinical practice guidelines already in place support this practice. **

**The evidence indicates that this measure is directly related to the process of care being delivered. **

1b. Performance Gap

Comments:

**YES. I actually thought the performance was very poor and there is a tremendous need for improvement (despite the NQF preliminary assessment of moderate). I would have preferred a larger dataset, but evidently this is the best they had. **

Data included was limited to 192 pts and the majority were white and female so it is difficult to know for sure if there is a performance gap and/or disparities in care. Clinics were either in CA, NJ or Maryland so no representation from the Midwest

**A gap has been documented both among clinics and even between top performing clinics and the ideal. Disparities have also been identified. **

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic Health Record, Paper Medical Records. This is not an eMeasure.

Specifications:

- The level of analysis is at the clinician-level.
- The numerator includes the number of patients who had an FN risk assessment documented in the medical record prior to the first cycle of intravenous chemotherapy. An FN risk assessment is defined as at least one of the following:
  - Template in the record or evidence that an online tool was used to assess FN risk (e.g., a Febrile Neutropenia Risk Assessment Tool similar to that described in the study by O'Brien et al. [2014])
  - FN risk of the planned regimen was noted as a percentage (e.g., >20%) OR noted qualitatively (e.g., "high FN risk")
  - Patient factor(s) was noted as a contributor to elevated FN risk (e.g., "high FN risk due to advanced age and comorbidity")
  - Justification for USE of CSF was documented (e.g., "high risk regimen, CSF support will be used;" "due to

the presence of expanders and risk of infection, CSF will be used")
    - Justification for NOT using CSF was documented (e.g., "due to patient's youth and excellent health, CSF support will not be used")
- The denominator includes the number of patients 18 years of age or older with a solid malignant tumor or lymphoma receiving the first cycle of intravenous chemotherapy. The denominator includes patients who meet the following conditions:
    - Patient was 18 years of age or older when first-cycle intravenous chemotherapy of the current regimen was initiated.
    - Patient's first-cycle intravenous chemotherapy was initiated any time during months 2 through 12 of the 12-month measurement period.
    - The treatment ordered was intravenous chemotherapy (see sheet labeled "IV Chemotherapy" in the attached Excel file for a list of CPT procedure codes for chemotherapy).
    - Patient was being treated for a solid malignant tumor or lymphoma (see sheets labeled "Denom Diagnoses ICD9," "Denom Diagnoses ICD10," and "Denom Diagnoses ICD9-ICD10" in the attached Excel file for a list of ICD-9-CM diagnosis codes, ICD-10 CM diagnosis codes, and a conversion table between ICD-9-CM and ICD-10-CM diagnosis codes, respectively).
    - Patient did not receive chemotherapy in the 12 months prior to the first cycle of chemotherapy.
    - Patients receiving experimental therapy or participating in clinical trials are not eligible because the trial protocol dictates CSF prophylaxis decisions.
    - Patients on weekly chemotherapy regimens are not eligible because the intervals between treatments are not long enough for CSF prophylaxis to have an effect.
    - Patients receiving concurrent radiation therapy (see sheet labeled "Radiation Therapy" in the attached Excel file for CPT codes) are not eligible because CSF prophylaxis is contraindicated for those patients due to the risk of irreversible stem cell damage. Patients who receive palliative local radiation for pain control are eligible.
    - Record of care was complete (e.g., provider notes prior to cycle #1 of chemotherapy are available).
- There are no denominator exclusions.
- The ICD-9, ICD-10, and CPT codes have been included in the specification details.
- The measure is not risk-adjusted.
- The measure results may be stratified by:
    - Age – Divided into five categories: 18-44, 45-64, 65-74, 75-84, and 85+ years
    - Race/Ethnicity
    - Gender
    - Curative/adjuvant and palliative chemotherapy
    - Periodicity of chemotherapy (2-, 3- and 4-week cycles)
- A calculation algorithm is provided and describes the process of calculating the measure.
- The measure data collection tool is provided with this submission form. The field test data collection form is available from the developer upon request.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing attachment

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.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset included a total of **192** patient records from 5 community oncology clinics from April 2011 through February 2016. Inclusion critera provided to the clinics for 40 randomly selected patient records included: age at least 18 years, solid tumor or lymphoma, initiating chemotherapy, and not participating in a clinical trial. A total of 200 records were received from the clinics but six patients were not eligible because of incomplete records, one due to malignancy other than solid tumor or lymphoma, and one due to concurrent radation.
    - The four most frequent cancers in the sample were breast, lymphoma, lung, and colon. The developer provided the characteristics of the clinics and demographic and clinical characteristics of the patients.
- Inter-rater reliability testing was assessed using two abstractors who were instructed to abstract the same randomly selected 50 medical records, 10 records per clinic, for a 25 percent inter-rater reliability (IRR) sample. The kappa statistic and percent agreement between the abstractors was calculated based on whether documentation of a febrile neutropenia risk assessment was in the medical record; this is an appropriate method for assessing patient-level data elements.
    - Kappa values range between 0 and 1. 0 and are interpreted as degree of agreement beyond chance. By convention, a kappa > .70 is considered acceptable inter-rater reliability.
        - 0 No better than chance
        - 0.01-0.20 Slight
        - 0.21-0.40 Fair
        - 0.41-0.60 Moderate
        - 0.61-0.80 Substantial
        - 0.81-1.0 Almost perfect

Results of reliability testing:

- The developer provided the kappa statistic and percent agreement between abstractors for scoring whether documentation of a febrile neutropenia risk assessment was in the medical records for each of the five clinics below.

Table 3. Inter-rater Reliability for Scoring Febrile Neutropenia Risk Assessment in Medical Record, by Clinic

Site	#of Medical Records in IRR Sample	Kappa Statistic (SE)	Agreement*
Clinic 1	10	1.0	100%
Clinic 2	10	0.783 (0.201)	90%
Clinic 3	10	1.0	100%
Clinic 4	10	1.0	100%
Clinic 5	10	1.0	100%
All sites	50	0.878 (0.120)	98%

*Percent of records for which the two coders agreed as to whether there was or was not a documented reference to FN risk in the record.

- For the inter-rater reliability sample, kappa estimates ranged from **0.783** to **1.0** for the five clinics.
- The developer provided kappa statistics and percent agreement results for one data element included in the numerator (documentation of a febrile neutropenia risk assessment in the medical record). NQF guidance states that testing should be done for all critical data elements.
- The clinics determined which patients met the denominator inclusion criteria (age at least 18 years, solid tumor or lymphoma, initiating chemotherapy, and not participating in a clinical trial). The developers excluded additional patients due to incomplete records, malignancy other than solid tumor or lymphoma, or concurrent radation.

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) → Computed performance scores for measure entities (Box 4) →Patient-level data elements (Box 8) →Appropriate for

assessing the reliability of critical data elements (Box 9) → High or moderate certainty or confidence (Box 10a) → Moderate (highest eligible rating is MODERATE).

Question for the Committee:

- *Does the measure consistently identify and include patients 18 years of age or older with a solid malignant tumor or lymphoma receiving the first cycle of intravenous chemotherapy?*
- *Is the test sample adequate to generalize for widespread implementation?*

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

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer assessed face validity using an panel of 10 experts in clinical oncology. The expert panel was asked to review the measure specifications and the evidence supporting the measure and determine if performance socres resulting from the measure as defined can be used to distinguish good from poor quality.

Validity testing results:

- **80%** (8/10) of the respondents either agreed or strongly agreed that performance scores resulting from the measure as defined can be used to distinguish good and poor quality.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

There are no exclusions in this measure.

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provided the measure rates by clinic and statistically significant differences between clinics in the tables below.

Table 7. Numerator, Denominator, and Measure Rate by Clinic

Clinic Number	Numerator	Denominator	Measure Rate
Clinic 1	6	37	0.162
Clinic 2	10	37	0.270
Clinic 3	7	39	0.179
Clinic 4	0	39	0.000
Clinic 5	1	40	0.025

Table 8. Statistical Significance of Comparisons Between Clinics Based on Two-Tailed Significance Test

	Clinic 1	Clinic 2	Clinic 3	Clinic 4
Clinic 1				
Clinic 2	NS*			
Clinic 3	NS*	NS*		
Clinic 4	P<0.05**	P<0.001**	P<0.05**	
Clinic 5	NS**	P<0.01**	P<0.05**	NS**

NS= Not significant at P<0.05
*Based on test for difference between two independent proportions.
**Based on a Fisher exact probability test.

Question for the Committee:
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- This was not performed for this measure.

2b7. Missing Data

- The developer stated that missing data was not identified udring the medical record abstraction.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1)→Threats to validity mostly assessed (Box 2) →Empirical validity testing (Box 3)→ Face validity assessed (Box 4)→ Agreement measure can be used to distinguish quality (Box 5)→ Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:
Again, would have preferred larger dataset, but appears reliable.
**Specifications are clearly defined. **
**The data elements are clearly defined, but somewhat complex. Once you have the data, the calculation logic must be carefully followed. Altogether may be a difficult measure to calculated consistently. **
Specifications are consistent with evidence

2a2. Reliability Testing

Comments:
**A relatively small sample was used for reliability testing, but it was successful. The developers excluded "additional patients due to incomplete records" but don't state how many patients were excluded. I would like to know about these incomplete records (how many? why were they incomplete? etc?) **
**Testing was not done for all critical data elements. **
**When carefully applied this measure does identify appropriate numerator and denominator values. It could be implemented generally, but would take some effort. **

2b2. Validity Testing

Comments:

**Small sample of experts for face validity is favorable. **

**A panel of 10 experts in clinical oncology were used and 80% of the respondents either agreed or strongly agreed that the scores could be used to distinguish good and poor quality. **

**It would be reasonable to draw conclusions about the thoroughness of care from this measure and it probably correlates with quality. **

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

**There are no exclusions and no risk adjustment. I would very much like to see data showing that groups with high scores on the measure have less FN, but unfortunately those data evidently don't exist. Missing data may constitute a threat (see response to 2a2). **

**Missing data is unlikely when the measure is used as described. If applied equally, then meaningful differences in quality may be evident. **

Criterion 3. Feasibility

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 electronic sources; information about FN risk assessment may require manual chart abstraction.
- Data collection burden due to manual chart abstraction requirement for FN risk assessment documentation.
- There are no fees to use the measure.

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

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

I suspect that EMRs can create automatic phrases to help improve the feasibility of the measure. However, currently there is a risk inherent in manual abstraction, and large volume practices may not be able to keep up.

Some data sources can be extracted from electronic sources, others may require manual chart abstraction.

Seems potentially very difficult to abstract the data, but could be done.

Criterion 4: Usability and Use

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:

- The developer stated that because the measure is being submitted to NQF for initial endorsement, they do not yet have specific plans to submit it for use in a specific federal, state or local program. However, the measure would be appropriate for use in a CMS reporting program for outpatient care provided to oncology patients, for example, under oncology bundled payment demonstrations. The developer will explore the possibility of submitting the measure to CMS for one of the reporting programs through the Measures under Consideration (MUC) process .
    - The developer did not provide expected timeframe/time line for submission of this measure to CMS through the MUC process or other implementation plan.

Improvement results:

- Progress on improvement is not required because this measure is being submitted for initial endorsement.

Unexpected findings (positive or negative) during implementation:

- N/A – This measure is not a maintenance measure.

Potential harms:

- The developer reports no unintended negative consequences were identified during testing.

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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

**Not yet [currently in use], but the plan is for future use in an accountability measure. **
**Not currently in use in any programs but could be part of a CMS reporting program for outpt care. **
**It's a satisfactory measure, but may not be high impact as it seems more daunting than others to operationalize. **

Criterion 5: Related and Competing Measures

Related or competing measures

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Assigned by NQF
Measure Title: Febrile Neutropenia Risk Assessment Prior to Chemotherapy
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: 3/11/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering 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:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
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

☐ Health 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

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

☒ Process: Febrile Neutropenia Risk Assessment Prior to Chemotherapy

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

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

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

Not applicable; the measure does not relate to a health outcome.

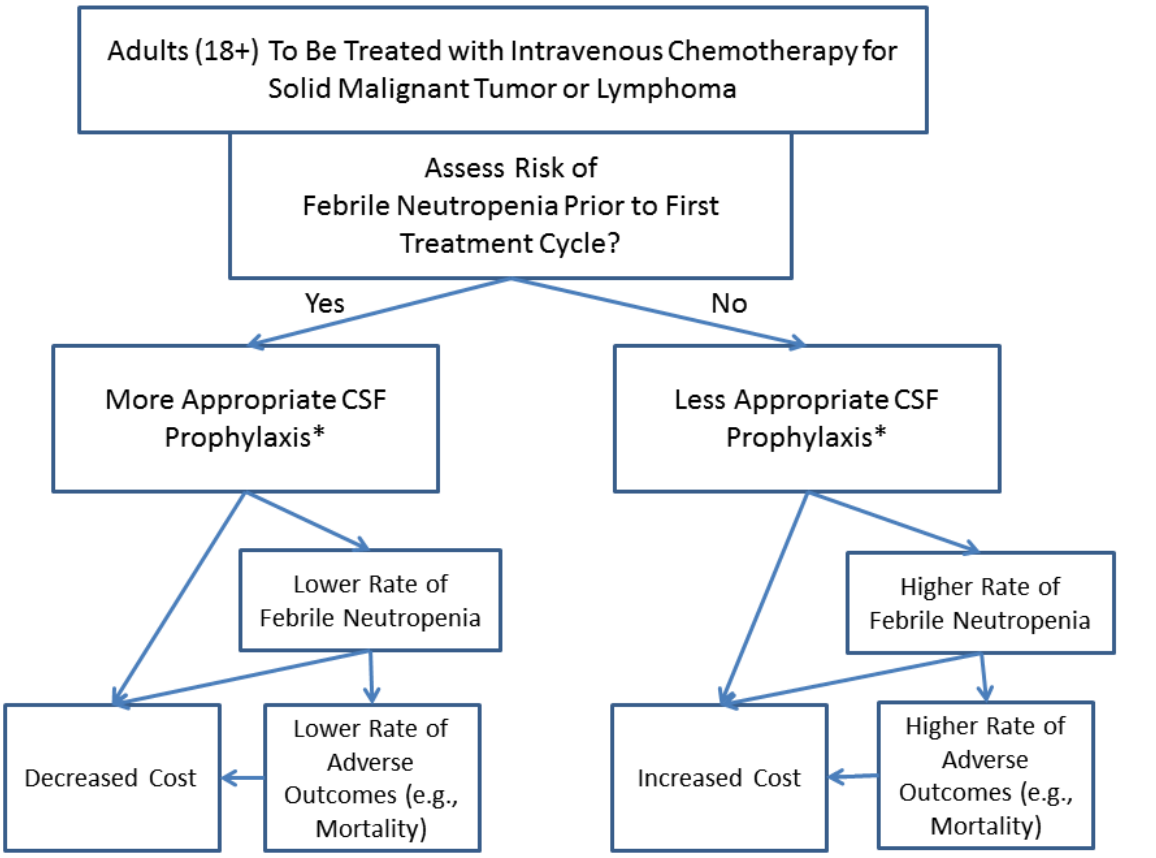
1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Not applicable; the measure does not relate to a health outcome.

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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



**Appropriate CSF prophylaxis means that patients who have an approximately 20% or higher risk for febrile neutropenia based on patient-, disease- and treatment-related factors receive primary prophylaxis and lower risk patients do not.*

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

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

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

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

☒ Other – ***complete section 1a.8***

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

Two clinical practice guidelines contain recommendations that support risk assessment for chemotherapy-induced febrile neutropenia (FN):

- 2015 American Society of Clinical Oncology (ASCO) Recommendations for the Use of WBC Growth Factors, and
- NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®).

The citation and URL are provided below for each of these guidelines:

2015 ASCO Clinical Practice Guideline Update:

Smith TJ, Bohlke K, Lyman GH, Carson KR, Crawford J, Cross SJ, Goldberg JM, Khatcheressian JL, Leighl NB, Perkins CL, Somlo G, Wade JL, Wozniak AJ, Armitage JO. Recommendations for the Use of WBC Growth Factors: American Society of

Clinical Oncology Clinical Practice Guideline Update. J Clin Oncol. 2015 October 1; 33(28): 3199–3212. Available December 2, 2015, at http://jco.ascopubs.org/content/33/28/3199.full.pdf+html

URL:
http://jco.ascopubs.org/content/33/28/3199.full.pdf+html

NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®)
Crawford, J., Becker, P. S., Armitage, J. O. et al. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®), Myeloid Growth Factors, Version 1.2015. Available December 2, 2015, at http://www.nccn.org/professionals/physician_gls/pdf/myeloid_growth.pdf

URL:
http://www.nccn.org/professionals/physician_gls/pdf/myeloid_growth.pdf

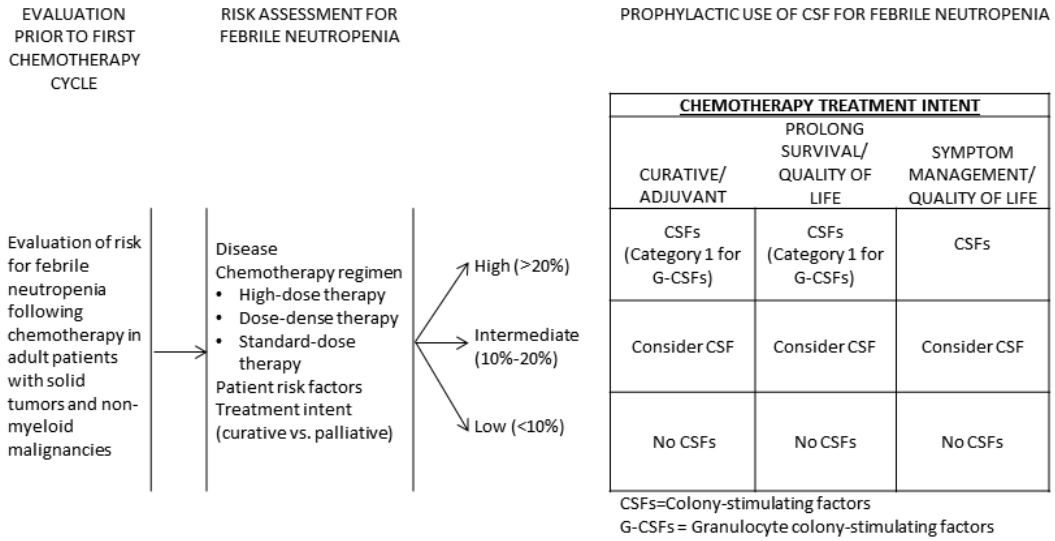
1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

The two guideline recommendations listed below provide strong support for conducting an FN risk assessment prior to the first cycle of chemotherapy to determine which patients with solid tumors should receive primary prophylaxis with a CSF. Bold italics were added in each recommendation to emphasize the text related to conducting an FN risk assessment.

Recommendation 1 (page 3203 in 2015 ASCO Clinical Practice Guideline Update; see complete citation in Section 1a.4.1 above):
"Recommendation 1: Primary prophylaxis with a CSF starting with the first cycle and continuing through subsequent cycles of chemotherapy is recommended in ***patients who have an approximately 20% or higher risk for febrile neutropenia based on patient-, disease- and treatment-related factors***. Primary CSF prophylaxis should also be administered in patients receiving dose dense chemotherapy when considered appropriate. Consideration should be given to alternative, equally effective, and safe chemotherapy regimens not requiring CSF support when available." (Emphasis added.)

Recommendation on Risk Assessment for Prophylactic Use of CSFs (page MS-9 in NCCN Clinical Practice Guidelines in Oncology [NCCN Guidelines®]; see complete citation in Section 1a.4.1 above):
"The guidelines begin with ***an evaluation of risk for chemotherapy-induced FN prior to the first cycle of chemotherapy***. The risk assessment includes disease type, chemotherapeutic regimen (high-dose, dose-dense, or standard-dose therapy), patient risk factors, and treatment intent. Three categories based on the intent of chemotherapy have been designated by the NCCN Panel. These include curative-adjuvant therapy, treatment directed toward prolongation of survival, and symptom management therapy. Based on the chemotherapy regimen and patient-related risk factors, the patient is assigned to either an overall high-risk group (>20% risk of FN), intermediate risk group (10%-20% risk), or low-risk group (<10% risk). Of note, there is currently no consensus nomogram for risk assessment. While the NCCN Panel outlines criteria to aid in the assessment of FN risk, independent clinical judgment should be exercised based on the patient's situation (see Patient Risk Factors for Developing Febrile Neutropenia in the algorithm). In addition to assessing patient- and treatment-related risk, consideration should be given to the intent of cancer treatment when determining the appropriate use of CSFs. For example, a patient with a previous neutropenic complication in the immediately prior cycle of chemotherapy, with no plan to reduce the dose intensity should be considered high risk." (Emphasis added.)

Diagram related to Recommendation on Risk Assessment for Prophylactic Use of CSFs (page MGF-1 in NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®); see complete citation in Section 1a.4.1 above):
The NCCN Algorithm shown on page MGF-1 of the 2015 NCCN Clinical Practice Guidelines on Myeloid Growth Factors (**NCCN Guidelines®**) and reproduced below illustrates the factors leading to the decision about prophylactic use of CSF for febrile neutropenia that is described in the paragraph above:



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

ASCO Recommendation 1:
Grades assigned to Recommendation 1 (page 3203 of 2015 ASCO Clinical Practice Guideline Update; see complete citation in Section 1a.4.1 above):
Type: **Evidence-based**, benefits outweigh harms.
Strength of recommendation: **Strong**.

Definitions:
Definition of "Evidence-based" Rating for Type of Recommendation (page 7 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 below):
Evidence-based="There was sufficient evidence from published studies to inform a recommendation to guide clinical practice."

Definition of "Strong" Rating for Strength of Recommendation (page 8 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 below):
Strong= "There is high confidence that the recommendation reflects best practice. This is based on (1) strong evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with no or minor exceptions; (3) minor or no concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation."

NCCN Recommendation on Risk Assessment for Prophylactic Use of CSFs from NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®):
Grade assigned to the quoted recommendation (pages MGF-1 and MS-9 from 2015 NCCN Clinical Practice Guidelines on Myeloid Growth Factors; see complete citation in Section 1a.4.1 above):
Category 2A (page MGF-1)

Definition of Category 2A (page MS-1 from 2015 NCCN Clinical Practice Guidelines on Myeloid Growth Factors [**NCCN Guidelines®**]; see complete citation in Section 1a.4.1 above):
Category 2A= "Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate."

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

2015 ASCO Clinical Practice Guideline Update:

Definitions of Other Types of Recommendation (page 7 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 below):
"Formal consensus: The available evidence was deemed insufficient to inform a recommendation to guide clinical practice. Therefore, the Expert Panel used a formal consensus process to reach this recommendation, which is considered the best current guidance for practice. The Panel may choose to provide a rating for the strength of the recommendation (i.e., "strong," "moderate," or "weak"). The results of the formal consensus process are summarized in the guideline and reported in the Data Supplement."

"Informal consensus: The available evidence was deemed insufficient to inform a recommendation to guide clinical practice. The recommendation is considered the best current guidance for practice, based on informal consensus of the Expert Panel. The Panel agreed that a formal consensus process was not necessary for reasons described in the literature review and discussion. The Panel may choose to provide a rating for the strength of the recommendation (i.e., "strong," "moderate," or "weak")."

"No recommendation: There is insufficient evidence, confidence, or agreement to provide a recommendation to guide clinical practice at this time. The Panel deemed the available evidence as insufficient and concluded it was unlikely that a formal consensus process would achieve the level of agreement needed for a recommendation."

Definitions of Other Ratings for Strength of Recommendation (page 8 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 below):

"Moderate: There is moderate confidence that the recommendation reflects best practice. This is based on (1) good evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with minor and/or few exceptions; (3) minor and/or few concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a moderate recommendation."

"Weak: There is some confidence that the recommendation offers the best current guidance for practice. This is based on (1) limited evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, but with important exceptions; (3) concerns about study quality; and/or (4) the extent of panelists' agreement. Other considerations (discussed in the guideline's literature review and analyses) may also warrant a weak recommendation."

NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®):

Other NCCN Categories of Evidence and Consensus (page MS-12015 of 2015 NCCN Clinical Practice Guidelines on Myeloid Growth Factors [**NCCN Guidelines®**]; see complete citation in Section 1a.4.1 above):

"Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate."

"Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate."
"Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate."

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**
Complete Citation for ASCO definitions in Sections 1a.4.3 and 1a.4.4:
ASCO Guideline Methodology Supplement. Recommendations for the Use of White Blood Cell Growth Factors: American Society of Clinical Oncology Clinical Practice Guideline Update. Available December 2, 2015, at http://www.instituteforquality.org/sites/instituteforquality.org/files/METHODOLOGY%20SUPPLEMENT%20WBCGF.pdf

URL:
http://www.instituteforquality.org/sites/instituteforquality.org/files/METHODOLOGY%20SUPPLEMENT%20WBCGF.pdf

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

☒ Yes → ***complete section 1a.7***

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

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):
Not applicable

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.
Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:
Not applicable

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)
Not applicable

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**
Not applicable

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE
1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Citation:
ASCO Guidelines Data Supplement. 2015. Recommendations for the Use of White Blood Cell Growth Factors: American Society of Clinical Oncology Clinical Practice Guideline Update. Available December 2, 2015, at http://www.instituteforquality.org/sites/instituteforquality.org/files/DATA%20SUPPLEMENT%20WBCGF.pdf

URL: http://www.instituteforquality.org/sites/instituteforquality.org/files/DATA%20SUPPLEMENT%20WBCGF.pdf

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

Citation:

ASCO Guidelines Methodology Supplement. 2015. Recommendations for the Use of White Blood Cell Growth Factors: American Society of Clinical Oncology Clinical Practice Guideline Update. Available December 2, 2015, at http://www.instituteforquality.org/sites/instituteforquality.org/files/METHODOLOGY%20SUPPLEMENT%20WBCGF.pdf

URL:
http://www.instituteforquality.org/sites/instituteforquality.org/files/METHODOLOGY%20SUPPLEMENT%20WBCGF.pdf

Complete section 1a.7

_____

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

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

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

2015 ASCO Clinical Practice Guideline Update: To evaluate the effect of colony-stimulating factors (CSFs) on clinical outcomes (e.g., febrile neutropenia, all-cause and infection-related mortality) in adults or children with a solid tumor or lymphoma treated with chemotherapy

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

Grade assigned to the evidence:
Evidence quality: high (page 3203 of 2015 ASCO Clinical Practice Guideline Update; see complete citation in Section 1a.4.1 above)

Definition of Rating for Strength of Evidence (page 9 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 above):
"High= High confidence that the available evidence reflects the true magnitude and direction of the net effect (i.e., balance of benefits v harms) and that further research is very unlikely to change either the magnitude or direction of this net effect."

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

Definition of All Other Ratings for Strength of Evidence (page 9 of 2015 ASCO Guideline Methodology Supplement; see complete citation in Section 1a.4.5 above):

"Intermediate: Moderate confidence that the available evidence reflects the true magnitude and direction of the net effect. Further research is unlikely to alter the direction of the net effect; however, it might alter the magnitude of the net effect."

"Low: Low confidence that the available evidence reflects the true magnitude and direction of the net effect. Further research may change either the magnitude and/or direction this net effect."

"Insufficient: Evidence is insufficient to discern the true magnitude and direction of the net effect. Further research may better inform the topic. The use of the consensus opinion of experts is reasonable to inform outcomes related to the topic."

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 1992-2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

Seven meta-analyses, two RCTs, and one systematic review are included in the body of evidence cited as support for Recommendation 1 in the 2015 Recommendations in the Use of WBC Growth Factors: American Society of Clinical Oncology Clinical Practice Guideline Update (see Section 1a.4.1 in this form for complete citation of this Guideline). One meta-analysis, three clinical practice guidelines, one RCT, and one systematic review which were cited by the ASCO Guidelines were excluded from this Measure Submission Form because they do not provide specific evidence to support the measure topic.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

For information related to the "overall quality of evidence across studies", see "Appendix Table 2. Quality of Methods Used in Studies Cited in Support of Recommendation 1 by ASCO Guidelines on the Use of WBC Growth Factors (Smith et al., 2015)" in the Appendix below.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

Excerpt from page 3203 of the 2015 ASCO Clinical Practice Guideline Update (see complete citation in Section 1a.4.1 above):

"Of the 16 publications that addressed primary prophylaxis (eight meta-analyses, three clinical practice guidelines, three RCTs, and two systematic reviews), none prompted a change in the level of febrile neutropenia risk warranting primary prophylaxis with a CSF. … ***The 20% cutoff for febrile neutropenia risk has been maintained from the 2005 guideline based on the evidence from randomized trials, especially the trial of CSFs in patients with breast cancer*** (Vogel et al., 2005), in which the baseline risk for febrile neutropenia was 17%. Independent systematic reviews of eight trials with 2,156 patients with breast cancer confirmed that CSFs reduce the risk of febrile neutropenia, with possible reductions in the need for hospitalization and all-cause mortality, but with no effect on infection-related mortality (Renner et al., 2012). Subsequent studies have shown that CSFs can reduce the risk of hospitalization for febrile neutropenia in elderly patients (age > 65 years) with solid tumors from 9% in all cycles to 5% (Balducci et al., 2007), but ***no other differences, such as in mortality, have been reported to justify treating a large number of patients who would not benefit and would experience potential toxicities and costs***." (Emphasis added.)

"However, recent publications have provided additional information about the likely benefits of primary prophylaxis. Meta-analyses of RCTs conducted in varying patient populations have confirmed that primary prophylaxis with a CSF reduces the risk of febrile neutropenia during chemotherapy for a solid tumor or lymphoma (Bohlius, Herbst, Reiser, Schwarzer, & Engert, 2008; Cooper, Madan, Whyte, Stevenson, & Akehurst, 2011; Kuderer, 2011; Kuderer, Dale, Crawford, & Lyman, 2007; Renner et al., 2012; Sung, Nathan, Alibhai, Tomlinson, & Beyene, 2007). Primary prophylaxis may also reduce the risk of hospitalization (Renner et al., 2012) and infection (Bohlius et al., 2008; Sung et al., 2007). Results for all-cause or infection-related mortality are less consistent. A meta-analysis of 59 RCTs among patients with solid tumors or lymphoma reported that primary prophylaxis with a G-CSF was associated with a modest reduction in all-cause mortality compared with no primary prophylaxis (risk ratio [RR], 0.93; 95%CI, 0.90 to 0.96; absolute risk difference, -3.2%; 95% CI, -2.1% to -4.2%) (Lyman et al., 2013). The greatest benefit was observed among patients who received dose-dense chemotherapy. In studies that evaluated the same dose and schedule of chemotherapy in different treatment arms, primary prophylaxis did not have a statistically significant effect on mortality. (Lyman et al., 2013). Another large meta-analysis considered 148 RCTs of primary prophylaxis in children or adults who were receiving cancer chemotherapy or undergoing stem-cell transplantation (SCT) (Sung et al., 2007). Only RCTs in which all study arms received the same chemotherapy or SCT conditioning regimen were included. On the basis of the 80 trials with all-cause

mortality results, short-term all-cause mortality was 7.6% with primary prophylaxis and 8.0% without primary prophylaxis (RR, 0.95; 95% CI, 0.84 to 1.08). Results for infection-related mortality were also null (RR, 0.82; 95% CI, 0.66 to 1.02) (Sung et al., 2007). In contrast, the addition of a G-CSF was associated with a statistically significant reduction in infection-related mortality in a 2011 meta-analysis of 12 RCTs in adults with a solid tumor or lymphoma; risk was 1.5% among patients who received primary prophylaxis with a CSF, compared with 2.8% among patients who did not receive primary prophylaxis (RR, 0.55; 95% CI, 0.34 to 0.90) (Kuderer, 2011)."

Here we list studies published from 2006-2012 that were cited in the 2015 ASCO Clinical Practice Guideline Update and reported on the effects of CSF on outcomes for adults with solid tumors:

- Febrile neutropenia (Renner et al., 2012; Balducci et al., 2007; Kuderer et al., 2007; Sung et al., 2007)
- All-cause mortality (Lyman et al., 2013; Lyman et al., 2010; Bohlius et al., 2008; Sung et al., 2007)
- Infection-related mortality (Renner et al., 2012; Bohlius et al., 2008; Kuderer et al., 2007; Sung et al., 2007)
- Early mortality (Renner et al., 2012; Kuderer et al., 2007)
- Planned chemotherapy doses at scheduled times and doses (Renner et al., 2012; (Wildiers & Reiser, 2011); Balducci et al., 2007; (Papaldo et al., 2006)
- Acute myeloid leukemia/ myelodysplastic syndrome (AML/MDS) (Lyman et al., 2010)
- Infections (Bohlius et al., 2008; Sung et al., 2007)

For more detail about the results from these studies, see "Appendix Table 1. Summary of Studies Cited in Support of Recommendation 1 by ASCO Guidelines on the Use of WBC Growth Factors (Smith et al., 2015)" in the Appendix below under the column heading "Benefit of Prophylaxis with CSF."

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

Excerpt from page 3203 of the 2015 ASCO Clinical Practice Guideline Update (see complete citation in Section 1a.4.1 above):
"Adverse effects of CSFs include bone pain, but a randomized trial of naproxen versus placebo suggested that nonsteroidal anti-inflammatory drugs may reduce the incidence, duration, and severity of bone pain among CSF-treated patients. (Kirshner et al., 2012). Naproxen was administered at a dose of 500 mg twice per day starting on the day of pegfilgrastim administration and continuing for 5 to 8 days."

Other "harms" of prophylaxis with CSF were reported in studies cited by the 2015 ASCO Clinical Practice Guideline Update, including injection-site reactions, arthralgia, and anemia. For more detail on bone pain and the other conditions, see "Appendix Table 1. Summary of Studies Cited in Support of Recommendation 1 by ASCO Guidelines on the Use of WBC Growth Factors (Smith et al., 2015)" in the Appendix below under the column heading "Adverse Events Associated with CSF Prophylaxis."

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

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

Not applicable.

1a.8 OTHER SOURCE OF EVIDENCE

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

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

The study by O'Brien et al. (2014) was identified by an oncologist on the research team. The other articles were identified by a citation search on the O'Brien et al. (2014) article and by a manual search of cited references in various articles.

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

Seven articles published from 2006 to 2016 provide insights into the benefits of FN risk assessment:

- Donohue (2006): Among patients receiving chemotherapy, the rates CSF prophylaxis were higher in those who were managed with a Risk Assessment Tool, than those in a "control group" that received care without use of the tool in an earlier time period (72% versus 28%, respectively, p<0.001). Conversely, the rates of adverse outcomes were higher in the control group than in the Risk Assessment Tool Group, but not statistically significant: febrile neutropenia (14% versus 11%, respectively), treatment with IV antibiotics (28% versus 14%), hospitalizations secondary to febrile neutropenia (16% versus 11%), and chemotherapy dose reductions (10% versus 3%).
- Doyle (2006): In a pre-post intervention study of patients initiating chemotherapy or a new regimen, use of tool for assessing patient risk of FN lowered the rate of FN-related hospitalization by 78%, from 9.7% among 155 patients in FY04 to 2.1% among 189 patients in FY05 (P = 0.003).
- Miller (2006): In a study of an intervention with a computer-based risk assessment tool (CBRAT), the rate of documenting performance of an FN risk assessment was 13% before use of the CBRAT and 100% after its introduction (p<0.001).
- O'Brien et al. (2014): An intervention study in a hospital-based oncology unit used an FN risk assessment tool to decide which patients receiving chemotherapy to treat with CSF. Comparing the time periods before (N=233 patients) and after (N=226 patients) the tool was used, the incidence of FN was reduced by 52% (p=0.02).
- Krzemieniecki et al. (2014): A total of 1,347 patients with solid tumors were eligible for the study based on being scheduled for "myelotoxic" chemotherapy and having an "investigator-assessed FN risk" of ≥ 20%. The study found 45-80% of these patients, depending on the tumor site, did not receive G-CSF that was indicated by results of the FN risk assessment by the investigator and guideline recommendations.
- Freyer et al. (2015): In a study of 165 physicians and 944 patients, each physician rated FN risk for their own patients using factors they selected. Only 82% of patients with an FN risk at or above 20% based on the physician-assessed FN risk were scheduled to receive CSF indicating almost one of five patients would not receive G-CSF PP even though the patient's risk was rated higher than the threshold of 20%.
- Mäenpää et al. (2016): In a study of 690 breast cancer patients (stages I-III) receiving chemotherapy, a higher proportion of those with a high-risk regimen were given G-CSF primary prophylaxis than those with a lower-risk regimen (48% versus 22%). However, these results indicate that less than half of patients on a high-risk regimen received appropriate treatment with G-CSF.

The full abstracts for these articles are provided below:

Citation: Donohue RB. (2006). Development and Implementation of a Risk Assessment Tool for Chemotherapy-Induced Neutropenia ONCOLOGY NURSING FORUM –33(2), 347-352.

"Purpose/Objectives: To evaluate a tool developed and implemented to help practitioners assess the risk of chemotherapy-induced neutropenia (CIN) and its complications in patients with nonleukemia cancer types."
"Design: Retrospective survey of chart records."
"Setting: Community-based oncology practice."
"Sample: The medical records of 85 adult patients treated with new courses of chemotherapy, regardless of the cancer type or stage; 50 charts belonged to patients treated before the implementation of the tool and 35 to patients evaluated with the tool."
"Methods: A risk assessment tool for CIN that was developed using risk factors from published studies and national guidelines was implemented. Patients who were found to be at increased risk for CIN were given colony-stimulating factor (CSF) support starting with the first chemotherapy cycle. The effectiveness of the tool was evaluated by comparing clinical outcomes before and after the implementation of the risk assessment tool."
"Main Research Variables: Febrile neutropenia, IV antibiotic use, hospitalization for neutropenia, and chemotherapy dose reductions and delays."

"Findings: Chemotherapy dose delays, febrile neutropenia, treatment with IV antibiotics, and hospitalization for neutropenia occurred less frequently in patients assessed with the tool and managed with the algorithm for CSF use than in those who were not assessed."
"Conclusions: The Risk Assessment for Neutropenic Complications Tool is effective in helping practitioners determine which patients are at high risk for CIN and its complications.
Implications for Nursing: By using the tool to identify patients treated with chemotherapy who need growth factor support, nurses can help to reduce the incidence of neutropenia and its complications."

Citation: Doyle AM. (2006). Prechemotherapy Assessment of Neutropenic Risk. ONCOLOGY Nurse Edition 20(10), 32-39.

"ABSTRACT: Chemotherapy-induced febrile neutropenia (FN) predisposes patients to life-threatening infections and typically requires hospitalization. The goal was to investigate whether a risk assessment tool aligned with national guidelines could help identify patients at risk of FN and reduce FN-related hospitalizations. Beginning in October 2004, oncology nurses applied the new risk assessment tool to all patients initiating chemotherapy or a new regimen. Patients at risk for FN received prophylactic colony stimulating factor. Charts for 189 patients receiving chemotherapy in fiscal year 2005 (FY05) were compared with charts of 155 patients receiving chemotherapy in FY04, before the tool was implemented. The incidence of FN-related hospitalization declined by 78%, from 9.7% in FY04 to 2.1% in FY05 (P = .003). Total hospital days decreased from 117 to 24. Routine systematic evaluation by oncology nurses improves recognition of patients at risk of FN and substantially reduces FN-related hospitalization."

Citation: Miller K. (2010). Using a Computer-Based Risk Assessment Tool to Identify Risk for Chemotherapy-Induced Febrile Neutropenia. Clinical Journal of Oncology Nursing 14(1), 87-91.

"This article evaluates the feasibility of developing and implementing a computer-based risk assessment tool (CBRAT) for febrile neutropenia and determines whether it could improve documentation of risk assessment in patients starting myelosuppressive chemotherapy regimens. The CBRAT was designed using a template creator in a commercial electronic medical records system. The effectiveness of the CBRAT was evaluated by comparing medical records data of patients with one or more risk factor for febrile neutropenia who were given prophylactic granulocyte–colony-stimulating factor before and after implementation. CBRAT usage significantly increased the likelihood of documented febrile neutropenia risk assessment from 13% before implementation to 100% after implementation ($p < 0.001$). No significant changes occurred in febrile neutropenia incidence rates, dose reductions, or dose delays. In addition, healthcare providers quickly learned how to operate the CBRAT and used it routinely, significantly improving the number of patients with documented febrile neutropenia risk assessment. Implementation of a computer-based tool can help nurses follow evidence-based guidelines that recommend routine febrile neutropenia risk assessment for patients initiating myelosuppressive chemotherapy."

Citation: O'Brien C, Dempsey O, Kennedy MJ. Febrile neutropenia risk assessment tool: Improving clinical outcomes for oncology patients. Eur J Oncol Nurs. 18 (2014) 167-174

"Purpose: To develop, implement and evaluate the effectiveness of a nurse-led risk assessment tool to reduce the incidence of febrile neutropenia (FN) and evaluate the nurse's role in FN risk assessment in a hospital-based oncology unit."
"Methods and sample: ***A FN risk assessment tool was developed, implemented and evaluated.*** A comparative prospective observational chart review was undertaken to evaluate the tool. Clinical data were collected from 459 patients' records from August 2008 through July 2009. Patients had no intervention during the first six months (n = 233). Patients in the following six months (n = 226) had the FN risk assessment completed and appropriate granulocyte-colony stimulating factor prescribed. A self-questionnaire was utilised to evaluate the nurses' role in FN risk assessment."
"Key results: ***The incidence of FN was reduced by 52% (p=0.02). Hospital days, dose reductions and treatment delays were reduced.*** Nurses felt they were the most appropriate person to carry out the assessment."

"Conclusions: Through consistent risk assessment, nurses could determine which patients were at high risk of developing FN leading to significant reduction in life-threatening infections, hospitalisations, dose reductions and delays. Nurses can be confident and competent in decision-making to reduce life threatening infections through the use of an FN risk assessment tool."
(Emphasis added.)

Citation: Krzemieniecki K, Sevelda P, Erdkamp F, Smakal M, Schwenkglenks M, Puertas J, et al. (2014). Neutropenia management and granulocyte colony-stimulating factor use in patients with solid tumours receiving myelotoxic chemotherapy - findings from clinical practice. Support Care Cancer 22, 667e77.

"Purpose: Clinical practice adherence to current guidelines that recommend primary prophylaxis (PP) with granulocyte colony stimulating factors (G-CSFs) for patients at high (≥20 %) overall risk of febrile neutropenia (FN) was evaluated."
"Methods: Adult patients with breast cancer, non-small cell lung cancer (NSCLC), small-cell lung cancer (SCLC), or ovarian cancer were enrolled if myelotoxic chemotherapy was planned, and they had an investigator-assessed overall FN risk ≥20 %. The primary outcome was FN incidence."
"Results: In total, 1,347 patients were analyzed (breast cancer, n =829; NSCLC, n =224; SCLC, n =137; ovarian cancer, n =157). ***Patients with breast cancer exhibited fewer individual FN risk factors than patients with other cancers and were far more likely to have received a high-FN-risk chemotherapy regimen. However, a substantial proportion of all patients (45–80 % across tumour types) did not receive G-CSF PP in alignment with investigator risk assessment and guideline recommendations.*** FN occurred in 127 patients overall (9%, 95% confidence interval (CI) 8–11%), and incidence was higher in SCLC (15%) than other tumor types (8% in ovarian and NSCLC, 9% in breast cancer). A post hoc analysis of G-CSF use indicated that G-CSF prophylaxis was not given within the recommended timeframe after chemotherapy (within 1–3 days) or was not continued across all cycles in 39% of patients."
"Conclusions: ***FN risk assessment was predominantly based on clinical judgement and individual risk factors, and guidelines regarding G-CSF PP for patients at high FN risk were not consistently followed. Improved education of physicians may enable more fully informed neutropenia management in patients with solid tumours."***
(Emphasis added.)

Citation: Freyer G, Kalinka-Warzocha E, Syrigos K et al. (2015). Attitudes of physicians toward assessing risk and using granulocyte colony-stimulating factor as primary prophylaxis in patients receiving chemotherapy associated with an intermediate risk of febrile neutropenia. Med Oncol 32, 236.

Abstract
"Febrile neutropenia (FN) is a potentially fatal complication of chemotherapy. This prospective, observational study describes physicians' approaches toward assessing FN risk in patients receiving chemotherapy regimens with an intermediate (10–20 %) FN risk. In the baseline investigator assessment, physicians selected factors considered important when assessing overall FN risk and deciding on granulocyte colony-stimulating factor (G-CSF) primary prophylaxis (PP). Physicians then completed patient assessments using the same lists of factors. The final FN risk scores and whether G-CSF PP was planned were reported. The final analysis included 165 physicians and 944 patients. The most frequently considered factor in both assessments was chemotherapy agents in the backbone (88 % of investigator and 93% of patient assessments). History of FN (83%), baseline laboratory values (76%) and age (73%) were commonly selected at baseline, whereas tumor type (72%), guidelines (62%) and tumor stage (43 %) were selected most during patient assessments. Median investigator-reported FN risk threshold for G-CSF PP was 20% (range 10–85***%). G-CSF PP was planned in 82%of patients with an FN risk at or above this threshold; therefore, almost one-fifth of qualifying patients would not receive G-CSF PP.*** Physicians generally follow guidelines, but also consider individual patient characteristics when assessing FN risk and deciding on G-CSF PP. ***A standardized FN risk assessment may optimize the use of G-CSF PP, which may minimize the incidence of FN in patients undergoing chemotherapy with an intermediate FN risk."***
(Emphasis added.)

Citation: Mäenpää J, Varthalitis I, Erdkamp F, Trojan A, Krzemieniecki K, Lindman H, et al. (2016). The use of granulocyte colony stimulating factor (G-CSF) and management of chemotherapy delivery during adjuvant treatment for early-stage breast cancer. Further observations from the IMPACT solid study. The Breast 25, 27e33

"Objective: To investigate the use and impact of granulocyte colony-stimulating factors (G-CSF) on chemotherapy delivery and neutropenia management in breast cancer in a clinical practice setting."
"Methods: IMPACT Solid was an international, prospective observational study in patients with a physician-assessed febrile neutropenia (FN) risk of ≥ 20%. This analysis focused on stages I-III breast cancer patients who received a standard chemotherapy regimen for which the FN risk was published. Chemotherapy delivery and neutropenia-related outcomes were reported according to the FN risk of the regimen and intent of G-CSF use."
"Results: 690 patients received a standard chemotherapy regimen; 483 received the textbook dose/ schedule with a majority of these regimens (84%) having a FN risk ≥ 10%. Patients receiving a regimen with a FN risk ≥ 10% were younger with better performance status than those receiving a regimen with a FN risk <10%. ***Patients who received higher-risk regimens were more likely to receive G-CSF primary prophylaxis (48% vs 22%), complete their planned chemotherapy (97% vs 88%) and achieve relative dose intensity ≥ 85% (93% vs 86%) than those receiving lower-risk regimens.*** Most first FN events (56%) occurred in cycles not supported with G-CSF primary prophylaxis."
"Conclusion: Physicians generally recommend standard adjuvant chemotherapy regimens and were more likely to follow G-CSF guidelines for younger, good performance status patients in the curative setting, and often modify standard regimens in more compromised patients. However***, G-CSF support is not optimal, indicated by G-CSF primary prophylaxis use in <50% of high-risk patients and observation of FN without G-CSF support."***
(Emphasis added.)

References

1. Balducci, L., Al-Halawani, H., Charu, V., Tam, J., Shahin, S., Dreiling, L., & Ershler, W. B. (2007). Elderly Cancer Patients Receiving Chemotherapy Benefit from First-Cycle Pegfilgrastim. *The Oncologist, 12*(12), 1416-1424.
2. Bohlius, J., Herbst, C., Reiser, M., Schwarzer, G., & Engert, A. (2008). Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. *Cochrane Database Syst Rev*(4), Cd003189.
3. Cooper, K., Madan, J., Whyte, S., Stevenson, M., & Akehurst, R. (2011). Granulocyte colony-stimulating factors for febrile neutropenia prophylaxis following chemotherapy: systematic review and meta-analysis. *BMC Cancer, 11*(1), 404.
4. Donohue, R. (2006). Development and implementation of a risk assessment tool for chemotherapy-induced neutropenia. *Oncol Nurs Forum, 33*(2), 347-352.
5. Doyle, A. M. (2006). Prechemotherapy assessment of neutropenic risk. *Oncology (Williston Park), 20*(10 Suppl Nurse Ed), 32-39; discussion 39-40.
6. Freyer, G., Kalinka-Warzocha, E., Syrigos, K., Marinca, M., Tonini, G., Ng, S. L., et al. (2015). Attitudes of physicians toward assessing risk and using granulocyte colony-stimulating factor as primary prophylaxis in patients receiving chemotherapy associated with an intermediate risk of febrile neutropenia. *Med Oncol, 32*(10), 236.
7. Kirshner, J. J., Heckler, C. E., Janelsins, M. C., Dakhil, S. R., Hopkins, J. O., Coles, C., & Morrow, G. R. (2012). Prevention of Pegfilgrastim-Induced Bone Pain: A Phase III Double-Blind Placebo-Controlled Randomized Clinical Trial of the University of Rochester Cancer Center Clinical Community Oncology Program Research Base. *Journal of Clinical Oncology, 30*(16), 1974-1979.
8. Krzemieniecki, K., Sevelda, P., Erdkamp, F., Smakal, M., Schwenkglenks, M., Puertas, J., et al. (2014). Neutropenia management and granulocyte colony-stimulating factor use in patients with solid tumours receiving myelotoxic chemotherapy--findings from clinical practice. *Support Care Cancer, 22*(3), 667-677.
9. Kuderer, N. M. (2011). Meta-analysis of randomized controlled trials of granulocyte colony-stimulating factor prophylaxis in adult cancer patients receiving chemotherapy. *Cancer Treatment and Research, 157*, 127-143.
10. Kuderer, N. M., Dale, D. C., Crawford, J., & Lyman, G. H. (2007). Impact of Primary Prophylaxis With Granulocyte Colony-Stimulating Factor on Febrile Neutropenia and Mortality in Adult Cancer Patients Receiving Chemotherapy: A Systematic Review. *Journal of Clinical Oncology, 25*(21), 3158-3167.
11. Lyman, G. H., Dale, D. C., Culakova, E., Poniewierski, M. S., Wolff, D. A., Kuderer, N. M., et al. (2013). The impact of the granulocyte colony-stimulating factor on chemotherapy dose intensity and cancer survival: a systematic review and meta-analysis of randomized controlled trials. *Annals of Oncology, 24*(10), 2475-2484.
12. Lyman, G. H., Dale, D. C., Wolff, D. A., Culakova, E., Poniewierski, M. S., Kuderer, N. M., & Crawford, J. (2010). Acute Myeloid Leukemia or Myelodysplastic Syndrome in Randomized Controlled Clinical Trials of Cancer Chemotherapy With Granulocyte Colony-Stimulating Factor: A Systematic Review. *Journal of Clinical Oncology, 28*(17), 2914-2924.
13. Maenpaa, J., Varthalitis, I., Erdkamp, F., Trojan, A., Krzemieniecki, K., Lindman, H., et al. (2016). The use of granulocyte colony stimulating factor (G-CSF) and management of chemotherapy delivery during adjuvant treatment for early-stage breast cancer-Further observations from the IMPACT solid study. *Breast, 25*, 27-33.
14. Miller, K. (2010). Using a computer-based risk assessment tool to identify risk for chemotherapy-induced febrile neutropenia. *Clin J Oncol Nurs, 14*(1), 87-91.
15. O'Brien, C., Dempsey, O., & Kennedy, M. J. (2014). Febrile neutropenia risk assessment tool: improving clinical outcomes for oncology patients. *Eur J Oncol Nurs, 18*(2), 167-174.
16. Papaldo, P., Ferretti, G., Di Cosimo, S., Giannarelli, D., Marolla, P., Lopez, M., et al. (2006). Does Granulocyte Colony-Stimulating Factor Worsen Anemia in Early Breast Cancer Patients Treated With Epirubicin and Cyclophosphamide? *Journal of Clinical Oncology, 24*(19), 3048-3055.
17. Renner, P., Milazzo, S., Liu, J. P., Zwahlen, M., Birkmann, J., & Horneber, M. (2012). Primary prophylactic colony-stimulating factors for the prevention of chemotherapy-induced febrile neutropenia in breast cancer patients. *Cochrane Database Syst Rev, 10*, Cd007913.
18. Sung, L., Nathan, P. C., Alibhai, S. M. H., Tomlinson, G. A., & Beyene, J. (2007). Meta-analysis: Effect of Prophylactic Hematopoietic Colony-Stimulating Factors on Mortality and Outcomes of Infection. *Annals of Internal Medicine, 147*(6), 400-411.

19. Vogel, C. L., Wojtukiewicz, M. Z., Carroll, R. R., Tjulandin, S. A., Barajas-Figueroa, L. J., Wiens, B. L., et al. (2005). First and subsequent cycle use of pegfilgrastim prevents febrile neutropenia in patients with breast cancer: a multicenter, double-blind, placebo-controlled phase III study. *J Clin Oncol, 23*(6), 1178-1184.
20. Wildiers, H., & Reiser, M. (2011). Relative dose intensity of chemotherapy and its impact on outcomes in patients with early breast cancer or aggressive lymphoma. *Critical Reviews in Oncology/Hematology, 77*(3), 221-240.

Appendix

Summary of Additional Evidence in Support of Proposed Measure on Febrile Neutropenia Risk Assessment Prior to Chemotherapy

Appendix Tables 1 and 2 below summarize information about the studies cited by the ASCO Guidelines on the Use of WBC Growth Factors (Smith et al., 2015) in support of Recommendation 1. The set of studies cited by these ASCO guidelines included adult or pediatric patients and patients with solid tumors or lymphoma. To the extent possible, we dropped entire studies or subsets of results that were focused on pediatric patients to more accurately reflect the focus of the measure, which is on adult patients with a solid tumor or lymphoma. However, for a few studies, dropping the subset of results on children was not feasible, and we present results that include this subgroup (as noted in the last column of Table 1).

Appendix Table 1. Summary of Studies Cited in Support of Recommendation 1 by ASCO Guidelines on the Use of WBC Growth Factors
(Smith et al., 2015)

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
Lyman 2013 (1)	Meta-analysis of data from 61 randomized comparisons; outcome=all-cause mortality.	**From Table 1**: Patients with solid tumors Relative risk (RR) for G-CSF vs. no G-CSF for all-cause mortality Breast (N=20) RR=0.954 (CI 0.898, 1.013) Genitourinary (N=7) RR=0.946 (CI 0.884, 1.013) Lung (N=16) RR=0.930 (CI 0.882, 0.980) Lymphoma (N=16) RR=0.895 (CI 0.841, 0.952) Other (N=2) RR=0.867 (CI 0.630, 1.193)	Adverse events were reported in all 58 of 59 RCTs (98%) and were systematically reported in 51 of 59 RCTs. However, specific information about adverse events was not included in the Lyman 2013 article.	
Renner 2012	Meta-analysis (Number of Main Comparison studies: 6 Febrile Neutropenia (FN), 8 Early Mortality, 8 Infection-related mortality; Number of Secondary Outcome studies: 4 Severe Neutropenia, 3 Infections, 4 Hospitalizations, 4 IV antibiotics, 4 Chemotherapy, 3 Bone Pain, 2 Injection-site Reaction)	**From "Summary of Findings for the Main Comparison" table (Page 4):** Patients with breast cancer in randomized control trials receiving primary prophylactic G-CSF/GM-CSF vs. no primary G-CSF/GM-CSF: Febrile Neutropenia (N=2073) Risk Ratio (RR)=0.27 (CI 0.11,0.70) Early Mortality (N=2143) RR=0.32 (CI 0.13,0.77) Infection-related mortality (N=2143) RR=0.14 (CI 0.02,1.29) From "Additional Summary of Findings" table (Page 21): Severe Neutropenia (Grade IV) (N=712) RR=0.44 (CI 0.17,1.18) Infections (N=210) RR=0.86 (CI 0.72,1.02) Hospitalization (N=1149) RR=0.14 (CI 0.06,0.3) IV Antibiotics (N=1568) RR=0.35 (CI	Patients with breast cancer in randomized control trials receiving primary prophylactic G-CSF/GM-CSF vs. no primary G-CSF/GM-CSF: Bone pain (N=388) RR=5.88 (CI 2.54,13.6) Injection-site Reaction (N=262) RR=3.59 (CI 2.33,5.53)	Patients with breast cancer

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
		0.22,0.55) Rate of patients who received the planned chemotherapy doses at scheduled times and doses (N=1588) RR=1.05 (CI 0.97,1.13)		
Cooper 2011	Meta-analysis 13 studies (5 Pegfilgrastim, 10 Filgrastim, 5 Lenograstim); outcome=FN incidence.	**From Figure 2:** Risk ratios for FN incidence (**From Section 4.1.1**) Patients with solid tumors or lymphoma for Pegfilgrastim vs. none in five studies: 0.08 (CI 0.03, 0.18) - 0.77 (CI 0.23, 2.60) Overall for Pegfilgrastim (N=2060): 0.30 (CI 0.14, 0.65) (**From Section 4.1.2**) Patients with solid tumors or lymphoma for Filgrastim vs. none in ten studies: 0.25 (CI 0.09, 0.72) - 0.82 (CI 0.64, 1.04) Overall for Filgrastim (N=2183): 0.57 (CI 0.48, 0.69) (**From Section 4.1.3**) Patients with solid tumors or lymphoma for Lenograstim vs. none in five studies: 0.34 (CI 0.15, 0.77) - 0.83 (CI 0.64, 1.08) Overall for Lenograstim (N=467): 0.62 (CI 0.44, 0.88) Overall for all treatments (N=4710): 0.51 (CI 0.41, 0.62)	None listed in Cooper 2011 article (one study includes patients below 18 years old)	
Kuderer 2011	Meta-analysis (same results were reported in Kuderer 2007)	See Kuderer 2007 below for results.	See Kuderer 2007 below for adverse events.	Results from same meta-analysis were reported in an earlier article (see Kuderer 2007 below).
Wildiers 2011	Systematic review 7 studies of breast cancer and 11 studies of lymphoma; outcome=relative dose intensity (RDI)	Impact of G-CSF on RDI in patients with breast cancer **From Table 7:** Four studies showed statistically significantly higher rates of achieving target rates of RDI with G-CSF treatment (compared to no G-CSF or no primary G-CSF), three studies do not show a difference. Primary G-CSF prophylaxis led to a statistically significant reduction in chance of receiving RDI <85% OR=0.733 (CI 0.61,0.88) p=0.001 Impact of G-CSF on RDI in patients with lymphoma **From Table 8:** Seven studies showed statistically	None listed in article	

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
		significantly higher rates of achieving target rates of RDI with G-CSF treatment (compared to no G-CSF or no primary G-CSF), three studies do not show a difference. Primary G-CSF prophylaxis led to a statistically significant reduction in chance of receiving RDI <85% OR=0.70 (CI 0.55,0.89)		
Lyman 2010	Meta-analysis outcomes= acute myeloid leukemia/ myelodysplastic syndrome (AML/MDS) and all-cause mortality; 23 studies for AML/MDS and 25 studies for all-cause mortality;	**From Table 2:** Risk ratio for AML/MDS among patients treated with G-CSF vs. no CSF Breast (7 studies): 1.811 (CI 0.897, 3.656) Endometrial (2 studies): 2.916 (CI 0.305, 27.872) Germ Cells (1 study): 0.336 (CI 0.014, 8.170) Hodgkin's Lymphoma (1 study): 2.013 (CI 0.820, 4.942) Non-Hodgkin's Lymphoma (8 studies): 2.732 (CI 0.804, 9.280) Lung (3 studies):0.956 (CI 0.101, 9.072) Urothelial (1 study): 0.963 (CI 0.061, 15.229) Risk ratio for all-cause mortality among patients treated with G-CSF vs. no CSF Breast (7 studies): 0.902 (CI 0.815, 0.998) Endometrial (2 studies): 0.945 (CI 0.874, 1.021) Germ Cells (1 study): 0.849 (CI 0.568, 1.269) Hodgkin's Lymphoma (1 study): 0.660 (CI 0.452, 0.963) Non-Hodgkin's Lymphoma (10 studies): 0.895 (CI 0.832, 0.963) Lung (3 studies): 0.945 (CI 0.875, 1.021) Urothelial (1 study) : 0.868 (CI 0.772, 0.977)	None listed in article	
Bohlius 2008	Meta-analysis; 13 RCTs (11 survival rate, 6 FFTF, 8 neutropenia, 5 febrile neutropenia ANC<1000, 3 febrile neutropenia ANC<500, 11 infection, 4 parental antibiotics treatment, 11 mortality during chemotherapy, 12	**From Figure 1:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis survival rate in 11 studies: OR=0.33 (CI 0.03, 3.27) – 2.04 (CI 0.55, 7.59) Overall survival rate OR=0.97 (CI 0.87 to 1.09) **From Figure 10:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis freedom from treatment failure (FFTF) in six studies: OR=0.96 (CI 0.28, 3.31) – 1.41 (CI 0.45, 4.41) Overall FFTF OR=1.11 (CI 0.91, 1.35)	From Figure 52: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for bone pain in nine studies: RR=1.30 (CI 0.47, 3.60) – 14.29 (CI 0.84, 242.02) Overall bone pain RR=3.57 (2.09, 6.12) From Figure 59: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for thrombosis	Adults with lymphoma

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
	infection related mortality, 13 complete response) Adverse events (9 bone pain, 5 thrombosis and related complications, 2 skin rash, 2 infection site reaction, 2 myalgia, 4 mucositis, 2 headache)	**From Figure 11:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from neutropenia in eight studies: RR=0.42 (CI 0.27, 0.66) – 1.48 (CI 0.57, 3.82) Overall neutropenia RR=0.67 (CI 0.60, 0.73) **From Figure 22:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from febrile neutropenia (ANC <1000) in five studies: RR=0.23 (CI 0.03, 1.75) – 1.09 (CI 0.48, 2.48) Overall febrile neutropenia (ANC<1000) RR=0.74 (CI 0.62, 0.89) **From Figure 29:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from febrile neutropenia (ANC <500) in three studies: RR=0.42 (CI 0.27, 0.66) – 0.67 (CI 0.48, 0.94) Overall febrile neutropenia (ANC<500) RR=0.59 (CI 0.48, 0.72) **From Figure 30:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from infection in 11 studies: RR=0.25 (CI 0.09, 0.72) – 1.33 (CI 0.46, 3.85) Overall infection RR=0.74 (CI 0.64 to 0.85) **From Figure 40:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from parental antibiotic treatment in four studies: RR=0.09 (CI 0.00, 1.51) – 1.21 (CI 0.80, 1.83) Parental antibiotic treatment RR=0.82 (CI 0.57, 1.18) **From Figure 41:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from mortality during chemotherapy in 11 studies: RR=0.31 (CI 0.01, 6.94) – 3.07 (CI 0.66, 14.37) Overall mortality during chemotherapy RR=0.93 (CI 0.60, 1.43) **From Figure 42:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis from infection-related mortality during chemotherapy in 12 studies: RR=0.21 (CI 0.02, 1.76) – 6.14 (CI 0.77, 48.87) Overall infection-related mortality RR=0.93 (CI 0.51, 1.71) **From Figure 43:** Patients treated with G-CSF and GM-CSF compared to no prophylaxis for complete response in 13 studies: RR=0.88 (CI 0.66, 1.17) – 3.50 (CI 0.40, 30.77)	and related complications in five studies: RR= 0.34 (CI 0.01, 8.14) – 4.76 (CI 0.24, 96.16) Overall thrombosis and related complications RR=1.29 (CI 0.56, 3.01) From Figure 60: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for skin rash in two studies: RR=4.33 (CI 1.04, 18.01) – 11.24 (CI 2.73, 46.25) Overall skin rash RR=7.69 (CI 2.84, 20.82) From Figure 61: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for injection site reaction in two studies: RR=6.52 (CI 2.91, 14.58) – 6.91 (CI 0.36, 131.75) Overall injection site reaction RR=6.55 (3.01, 14.25) From Figure 62: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for myalgia in two studies: RR=0.87 (CI 0.56, 1.37) – 2.60 (CI 0.29, 23.50) Overall myalgia RR=0.94 (CI 0.60, 1.45) From Figure 63: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for mucositis in four studies: RR=0.81 (CI 0.43, 1.54) – 1.33 (0.30, 5.84) Overall mucositis RR=0.95 (CI 0.64, 1.41) From Figure 64: Patients treated with G-CSF and GM-CSF compared to no prophylaxis for headache in two studies: RR=1.14 (CI 0.40, 3.26) – 2.19 (CI 1.38 – 3.49)	

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
		Overall complete tumor response RR=1.03 (CI 0.95, 1.10)		
Balducci 2007	RCT; treatment= outcomes= febrile neutropenia, grade 3 or 4 neutropenia, grade 4 neutropenia, chemotherapy delays, chemotherapy dose reductions, antibiotic use associated with neutropenia-related events	**Physician Discretion vs. Pegfilgrastim on all cycles in elderly patients with solid tumors (N=701):** **From Figure 2**, Lower incidence of Febrile Neutropenia 10.0% vs. 4.0% (p=0.001) **From Figure 4A**, Lower incidence of Grade 3 or 4 Neutropenia 80% vs. 30% (Significant) **From text:** Lower incidence of Grade 4 Neutropenia 58% vs. 22% (Significant) **From Figure 4A**, Lower rates of antibiotic use associated with neutropenia-related events 10% vs. 28% (Significant) **From Figure 4A**, Lower rates of hospitalization for febrile neutropenia and neutropenia 9% vs. 5% (Not significant, NS) **From Figure 4A**, Fewer chemotherapy delays 28% vs. 16% (Significant) **From Figure 4A**, Fewer chemotherapy dose reductions 14% vs. 7% (NS) **Physician Discretion vs. Pegfilgrastim on cycle 1 on elderly patients with solid tumors:** **From Figure 2**, Lower incidence of Febrile Neutropenia 7% vs. 3% (NS) **From text**: Lower incidence of Grade 3 or 4 Neutropenia 68% vs. 26% (Significant) **Physician Discretion vs. Pegfilgrastim on all cycles in elderly patients with NHL (N=151):** **From Figure 2**, Lower incidence of Febrile Neutropenia 37.0% vs. 15.0% (p=0.004) **From Figure 4B**, Lower incidence of Grade 3 or 4 Neutropenia 90% vs. 82% (Not Significant, NS) **From text**: Lower incidence of Grade 4 Neutropenia 86% vs. 75% (NS) **From Figure 4B**, Higher rates of antibiotic use associated with neutropenia-related events 53% vs. 55% (NS) **From Figure 4B**, Lower rates of hospitalization for febrile neutropenia and neutropenia 37% vs. 17% (NS) **From Figure 4B**, More chemotherapy delays 23% vs. 29% (NS) **From Figure 4B**, More chemotherapy dose reductions 8% vs. 16% (NS)	Greater than or equal to 5% of patients experienced pancytopenia, pneumonia, pyrexia, dehydration, and syncope in studies of solid tumors and lymphomas. Pegfilgrastim treatment was associated with arthralgia in studies of solid tumors and lymphomas (no point estimates listed). Pegfilgrastim treatment vs. Physician Discretion in patients with solid tumors: Higher incidence of bone pain 12% vs. 5% Pegfilgrastim treatment vs. Physician Discretion in patients with NHL: Higher incidence of bone pain 9% vs. 4%	Patients ≥65 yo with lymphoma, lung, breast, or ovarian cancer

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
		From text: Lower incidence of grade 4 neutropenia 86% vs. 75% (NS) **Physician Discretion vs. Pegfilgrastim on cycle 1 on elderly patients with NHL:** **From Figure 2**, Lower incidence of Febrile Neutropenia 25% vs. 7% (NS) **From text**: Lower incidence of Grade 3 or 4 Neutropenia 88% vs. 69% (NS)		
Kuderer 2007 (results of the same meta-analysis also reported in Kuderer 2011)	Meta-analysis 12-15 studies (7-9 Filgrastim, 4-5 Lenograstim, 1 Pegfilgrastim) 14 studies for bone or musculoskeletal pain; outcomes=infection-related mortality, early mortality, and febrile neutropenia	**From Figure 2: Infection-Related Mortality** Range of relative risks for patients with solid tumor or lymphoma for Filgrastim vs. none in seven studies: 0.328 (CI 0.035, 3.073) – 1.095 (CI 0.226, 5.293) Overall for Filgrastim: 0.529 (CI 0.304, 0.921) Range of relative risks for patients with solid tumor or lymphoma for Lenograstim vs. none in four studies: 0.650 (CI 0.112, 3.790) – 1.174 (CI 0.024, 56.861) Overall for Lenograstim: 0.829 (CI 0.257, 2.680) Relative risk for patients with solid tumor or lymphoma for Pegfilgrastim vs. none in one study: 0.201 (CI 0.010, 4.172) Overall for all treatments: 0.552 (CI 0.338, 0.902) **From Figure 3: Early Mortality** Range of relative risks for patients with solid tumor or lymphoma for Filgrastim vs. none in eight studies: 0.206 (CI 0.024, 1.732) – 1.427 (CI 0.435, 4.675) Overall for Filgrastim: 0.603 (CI 0.410, 0.887) Range of relative risks for patients with solid tumor or lymphoma for Lenograstim vs. none in four studies: 0.767 (CI 0.294, 2.002) – 1.174 (CI 0.024, 56.861) Overall for Lenograstim: 0.837 (CI 0.383, 1.833) Range of relative risks for patients with solid tumor or lymphoma for Pegfilgrastim vs. none in one study: 0.359 (CI 0.130, 0.988) Overall for all treatments: 0.599 (CI 0.433, 0.830) **From Figure 4: Febrile Neutropenia** Range of relative risks for patients with solid tumor or lymphoma for	Risk ratio for patients treated with G-CSF with bone or musculoskeletal pain (N=3029) RR=4.023 (CI 2.156,7.52). These results may include patients under 18 years of age.	Three studies include patients below 18 years old.

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
		Filgrastim vs. none in nine studies: 0.249 (CI 0.087, 0.716) – 0.816 (CI 0.641, 1.039) Overall for Filgrastim: 0.614 (CI 0.525, 0.718) Range of relative risks for patients with solid tumor or lymphoma for Lenograstim vs. none in five studies: 0.338 (CI 0.148, 0.770) – 0.829 (CI 0.636, 1.080) Overall for Lenograstim: 0.623 (CI 0.442, 0.879) Range of relative risks for patients with solid tumor or lymphoma for Pegfilgrastim vs. none in one study: 0.077 (CI 0.034, 0.175) Overall for all treatments: 0.538 (0.430, 0.673) From Table 2: Summary risk ratios for patients with solid tumors treated with G-CSF vs. none: **Febrile Neutropenia 0.44 (CI 0.30, 0.65), Early Mortality 0.55 (CI 0.37, 0.84),** Infection-Related Mortality 0.53 (CI 0.28, 1.02). Summary risk ratios for patients with lymphomas treated with G-CSF vs. none: **Febrile Neutropenia 0.71 (CI 0.59, 0.85)**, Early Mortality 0.69 (CI 0.40, 1.17)**,** Infection-Related Mortality 0.58 (CI 0.28, 1.23). **Bold** = statistically significant differences.		
Sung 2007	Meta-analysis In Appendix Figure 3: 26 lymphoma/ solid tumor studies In Appendix Figure 6: 30 lymphoma/ solid tumor studies In Appendix Table 3: Number of G-CSF studies: 45 all-cause mortality, 42 infection-related mortality, 41 documented infections, 29 microbiologically documented infection, 33 febrile neutropenia.	**From Appendix Figure 3:** All-cause mortality associated with CSF for lymphoma/solid tumor patients (n=4359) Risk ratio (RR)=0.91 (CI 0.64,1.28) **From Appendix Figure 6:** Infection-related mortality associated with CSF for lymphoma/solid tumor patients (N=4777) RR=0.70 (CI 0.47,1.05) **From Appendix Table 3:** For patients treated with G-CSF: All-cause mortality RR=0.98 (0.83, 1.15) Infection-related mortality RR=0.84 (CI 0.66, 1.06) Documented infections RR=0.83 (CI 0.76, 0.91) Microbiologically documented infection RR=0.85 (CI 0.76, 0.96) Febrile neutropenia RR=0.72 (CI 0.64, 0.81)	None listed in article	Adults and children with cancer Limitations: these data are not limited to adults 18+.

Author Year	Type of Study	Benefit of Prophylaxis with CSF	Adverse Events Associated with CSF Prophylaxis	Notes
	Number of GM-CSF studies: 34 all-cause mortality, 24 infection-related mortality, 17 documented infections, 12 microbiologically documented infection, 15 febrile neutropenia.	**From Appendix Table 3:** For patients treated with GM-CSF All-cause mortality RR=0.95 (0.84, 1.08) Infection-related mortality RR=0.82 (CI 0.49, 1.38) Documented infections RR=0.92 (CI 0.78, 1.07) Microbiologically documented infection RR=0.90 (CI 0.68, 1.19) Febrile neutropenia RR=0.88 (CI 0.75, 1.03)		
Papaldo 2006	RCT; outcomes=delayed cycles; dose reduction, dose intensity	Impact of G-CSF on all cycles on patients with early breast cancer (N=506): Decreased rate of delayed cycles for patients with G-CSF (10.0% vs. 3.6% p=0.00001) Decreased frequency of dose reduction for patients with G-CSF (3.6% vs. 1.4% p=0.002) No difference in dose intensity of adjuvant therapy between the G-CSF and control (98.1% vs. 95.5% in G-CSF and non–GCSF, respectively; p=0.17).	Impact of G-CSF on all cycles on anemia rate in patients with early breast cancer (N=506): Mean hemoglobin value from cycle 3 onwards was lower for each cycle for patients with G-CSF (p<0.0001). Increased rate of Grade 2 or worse anemia in patients with G-CSF (38.8% vs. 26.2% p=0.005).	Female breast cancer patients

Appendix Table 2. Quality of Methods Used in Studies Cited in Support of Recommendation 1 by ASCO Guidelines on the Use of WBC Growth Factors (Smith et al., 2015)

Information about Quality of Study Methods
Lyman et al., 2013: Meta-analysis; 61 separate randomized comparisons of chemotherapy with (N= 11,337 patients) or without (N= 13,456 patients) the initial use of G-CSFs. Methodologic flaws of the meta-analysis may include: may have missed relevant studies; subject to the weaknesses of ecologic studies; control patients may have received CSFs later in the study; little or no data on dose and length of G-CSF treatment; survival of patients based on original study data; minimum follow-up of two years to be eligible for meta-analysis; biased selection of patient sample to include those without comorbidities and "poor performance"; possible publication bias.
Renner et al., 2012: Meta-analysis; methodologic flaws of the meta-analysis may include small number of studies (N=8) and number of patients (N= 2156); heterogeneous disease stages and chemotherapy treatments; outcome definitions varied across studies; some studies were conducted before current recommendations for CSF were in place.
Cooper et al., 2011: Meta-analysis; 20 studies compared primary G-CSF prophylaxis with no primary G-CSF prophylaxis; number of patients in analyses by type of G-CSF ranged from 467 to 4710. Methodologic flaws of the meta-analysis may include heterogeneity across studies in patient groups (age, cancer type), chemotherapy regimen, and number and length of cycles.
Wildiers et al., 2011: Systematic review with only one reviewer evaluating the studies; seven studies ranged in size from 41 patients with G-CSF and 403 without, to a study of 5253 patients with G-CSF and 14645 patients without.
Lyman et al., 2010: Meta-analysis; 25 studies were included in the meta-analysis, for a total of 12,804 patients (6,058 patients randomly assigned to G-CSF and 6,746 controls). Methodologic flaws of the meta-analysis may include not identifying some relevant studies; control patients may have received CSFs later in the study; little or no data on dose and length of G-CSF treatment making an analysis of a dose-response relationship impossible; possible publication bias.
Bohlius et al., 2008: Meta-analysis; 13 randomized controlled trials were included with 2607 randomized patients; the method of

Information about Quality of Study Methods
allocation concealment was unknown for three studies, but concealment of allocation was adequate in the other studies; five studies were placebo-controlled; seven studies and one substudy were based on an intention-to-treat analysis; the rest were "based on full set analysis and excluded patients who did not meet the eligibility criteria, had major protocol violation or did not receive any study medication."
Kuderer et al., 2007: Meta-analysis; Infection-related mortality was reported as an outcome in 12 trials with 1,454 control patients and 1,463 patients receiving G-CSF; early mortality was an outcome in 13 trials with 3,122 patients; FN was an outcome in 15 trials with 3,182 patients. Methodologic flaws of the meta-analysis may include lack of inclusion of dose-intensification trials; based on aggregate data not on individual patient data; small sample sizes limit ability to analyze data on secondary outcomes such as infection-related mortality, early mortality, and RDI; insufficient statistical power to detect effects; possible under-reporting of FN, FN-related mortality, and cost.
Balducci et al., 2007: Randomized controlled trial (RCT) of 701 patients with solid tumors; the quality of this randomized controlled trial (RCT) was evaluated by Smith et al. (2015) as having an overall risk of bias of "intermediate" based on the following: The RCT had adequate randomization, sufficient sample size, similar groups, validated and reliable measures, and adequate follow-up. However, the RCT was not blinded, did not perform intent-to-treat analyses, and had significant conflicts of interest.
Sung et al., 2007: Meta-analysis; 4,359 patients (2,204 treated, 2,155 control) in mortality analysis; 4,777 patients (2,413 treated, 2,364 control) in analysis of infection-related mortality. A total of 148 studies were rated on the Jadad scale for study quality which assesses whether randomization was adequate, double-blinding was performed, and withdrawals and dropouts were described. The median Jadad score was 2 (range, 0 [lowest quality] to 5 [highest quality]), with substantial interrater agreement (weighted 0.73 [CI, 0.66 to 0.80]). Patients in all included studies were randomized to either CSF, or to placebo or no treatment, but patient and study characteristics were heterogeneous. All-cause mortality, the primary outcome, was defined without heterogeneity, as were most secondary outcomes, but not all outcomes were reported for every study, which may mean there was "selective reporting" and possible bias.
Papaldo et al., 2006: RCT; 506 patients randomly assigned to treatment with G-CSF or no G-CSF; the quality of this RCT was evaluated by Smith et al. (2015) as having an overall risk of bias of "high" based on the following: It was unclear whether the RCT had adequate randomization, sufficient sample size, similar groups, or blinding, and intent-to-treat analyses were not performed. However, the RCT did have validated and reliable measures, and adequate follow-up. In addition, the study may have had significant conflicts of interest.

Appendix References

Aapro MS, Bohlius J, Cameron DA, et al: 2010 update of EORTC guidelines for the use of granulocyte-colony stimulating factor to reduce the incidence of chemotherapy-induced febrile neutropenia in adult patients with lymphoproliferative disorders and solid tumours. Eur J Cancer 47:8-32, 2011

Balducci L, Al-Halawani H, Charu V, et al: Elderly cancer patients receiving chemotherapy benefit from first-cycle pegfilgrastim. Oncologist 12:1416-24, 2007

Bohlius J, Herbst C, Reiser M, et al: Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. Cochrane Database Syst Rev:CD003189, 2008

Cooper KL, Madan J, Whyte S, et al: Granulocyte colony-stimulating factors for febrile neutropenia prophylaxis following chemotherapy: systematic review and meta-analysis. BMC Cancer 11:404, 2011

Herbst C, Naumann F, Kruse EB, et al: Prophylactic antibiotics or G-CSF for the prevention of infections and improvement of survival in cancer patients undergoing chemotherapy. Cochrane Database Syst Rev:CD007107, 2009

Kirshner JJ, Heckler CE, Janelsins MC, et al: Prevention of pegfilgrastim-induced bone pain: a phase III double-blind placebo-controlled randomized clinical trial of the university of rochester cancer center clinical community oncology program research base. J Clin Oncol 30:1974-9, 2012

Kuderer NM: Meta-analysis of randomized controlled trials of granulocyte colony-stimulating factor prophylaxis in adult cancer patients receiving chemotherapy. Cancer Treat Res 157:127-43, 2011

Kuderer NM, Dale DC, Crawford J, et al: Impact of primary prophylaxis with granulocyte colony-stimulating factor on febrile neutropenia and mortality in adult cancer patients receiving chemotherapy: a systematic review. J Clin Oncol 25:3158-67, 2007

Lyman GH, Dale DC, Culakova E, et al: The impact of the granulocyte colony-stimulating factor on chemotherapy dose intensity and cancer survival: a systematic review and meta-analysis of randomized controlled trials. Ann Oncol 24:2475-84, 2013

Lyman GH, Dale DC, Wolff DA, et al: Acute myeloid leukemia or myelodysplastic syndrome in randomized controlled clinical trials of cancer chemotherapy with granulocyte colony-stimulating factor: a systematic review. J Clin Oncol 28:2914-24, 2010

Papaldo P, Ferretti G, Di Cosimo S, et al: Does granulocyte colony-stimulating factor worsen anemia in early breast cancer patients treated with epirubicin and cyclophosphamide? J Clin Oncol 24:3048-55, 2006

Renner P, Milazzo S, Liu JP, et al: Primary prophylactic colony-stimulating factors for the prevention of chemotherapy-induced febrile neutropenia in breast cancer patients. Cochrane Database Syst Rev 10:CD007913, 2012

Sung L, Nathan PC, Alibhai SM, et al: Meta-analysis: effect of prophylactic hematopoietic colony-stimulating factors on mortality and outcomes of infection. Ann Intern Med 147:400-11, 2007

Wildiers H, Reiser M: Relative dose intensity of chemotherapy and its impact on outcomes in patients with early breast cancer or aggressive lymphoma. Crit Rev Oncol Hematol 77:221-40, 2011

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

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

NQF_2930_Evidence_Form_3-11-16_To_NQF.pdf,NQF_2930_Evidence_Form_3-11-16_To_NQF.docx

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., the benefits or improvements in quality envisioned by use of this measure)*

This process measure focuses on assessing the risk of febrile neutropenia (FN) in patients with a solid malignant tumor or lymphoma prior to receiving their first cycle of an intravenous chemotherapy regimen. FN is a complication of chemotherapy that occurs as a result of chemotherapy-induced neutropenia, causing the patient to be highly susceptible to infection. FN after chemotherapy occurs frequently, with the incidence of FN among patients with solid tumors estimated to be 13.1-20.6% during their chemotherapy course and 3.1-7.4% in the first cycle (Weycker et al., 2015), and incidence among patients with lymphoma estimated at 36% (Bohlius et al., 2008). If a patient presents with fever during the neutropenic phase, antibiotic treatment (usually intravenous) and often hospital admission are required to control a likely infection and prevent the development of sepsis, and other complications, including death. Estimates of mortality for patients who were hospitalized for complications related to FN range from 7 to 20 percent among those with solid tumors, with higher rates among those with comorbidities (Kuderer et al., 2006; Elting et al., 1997; Schwenkglenks et al., 2006; Segal et al., 2008), and 9 percent among those with lymphoma (Kuderer et al., 2006).

Having information about a patient's FN risk allows the identification of patients at higher risk of FN who are more likely to benefit from treatment with prophylactic colony-stimulating factor (CSF) which stimulates the production of white blood cells and lowers the risk of FN and its complications. If a higher proportion of patients are assessed for FN risk, more of those with a higher FN risk would receive CSF and a lower proportion of patients would be expected to develop FN and its complications.

Citations

Bohlius, J., Herbst, C., Reiser, M., Schwarzer, G., & Engert, A. (2008). Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. Cochrane Database Syst Rev(4), Cd003189.

Elting, L. S., Rubenstein, E. B., Rolston, K. V., & Bodey, G. P. (1997). Outcomes of bacteremia in patients with cancer and neutropenia: observations from two decades of epidemiological and clinical trials. Clin Infect Dis, 25(2), 247-259.

Kuderer, N. M., Dale, D. C., Crawford, J., Cosler, L. E., & Lyman, G. H. (2006). Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. Cancer, 106(10), 2258-2266.

Schwenkglenks M., J. C., Constenla M., Leonard R.C. (2006). Neutropenic event risk and impaired chemotherapy delivery in six European audits of breast cancer treatment. Supportive Care Cancer, 14(9), 901-909.

Segal, B. H., Freifeld, A. G., Baden, L. R., Brown, A. E., Casper, C., Dubberke, E., et al. (2008). Prevention and treatment of cancer-related infections. J Natl Compr Canc Netw, 6(2), 122-174.

Weycker, D., Li, X., Edelsberg, J., Barron, R., Kartashov, A., Xu, H., & Lyman, G. H. (2014). Risk and Consequences of Chemotherapy-Induced Febrile Neutropenia in Patients With Metastatic Solid Tumors. Journal of Oncology Practice, 11(1), 47-54.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Table 1. Summary Statistics for Measure Rates by Clinic

No. of clinics / Mean / Median / Min / Max / STD / IQR / P10 / P25 / P50 / P75 / P90

5 / 0.127 / 0.162 / 0 / 0.270 / 0.113 / 0.154 / 0.01 / 0.025 / 0.162 / 0.179 / 0.234

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.

FN occurs frequently after chemotherapy, with the incidence among patients with solid tumors estimated to be 13.1-20.6% during their chemotherapy course and 3.1-7.4% in the first cycle (Weycker et al., 2015), and incidence among patients with lymphoma estimated at 36% (Bohlius et al., 2008). Prophylactic treatment with CSF reduces the risk of FN substantially. For example, in a systematic review, use of CSFs was shown to significantly lower the proportion of breast cancer patients who developed FN after chemotherapy compared to placebo or no treatment (RR 0.27; 95% CI 0.11 to 0.70) (Renner et al., 2012). Additional data for other solid tumor sites and lymphoma are provided in the Evidence Form for this measure. Assessing FN risk prior to initiation of chemotherapy allows identifying patients who should receive CSF prophylaxis, thereby reducing the incidence of FN. According to a recent study in patients with solid tumors or lymphoma (O´Brien, Dempsey, & Kennedy, 2014), the rate of FN decreased by 52% when a tool is used to estimate FN risk and those with higher risk were treated with G-CSF.

There is limited published data on the frequency of risk assessment for FN. A study was conducted at four offices of a community oncology practice to assess the effect of a computer-based risk assessment tool (CBRAT) for FN in patients starting myelosuppressive chemotherapy regimens for breast cancer or non–small cell lung cancer (Miller, 2010). Before the implementation of the CBRAT, 13 of 101 patients (13%) had documented risk assessments for febrile neutropenia. After the implementation of CBRAT, risk assessment increased to 100% ($p < 0.001$).

Citations

Bohlius, J., Herbst, C., Reiser, M., Schwarzer, G., & Engert, A. (2008). Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. Cochrane Database Syst Rev(4), Cd003189.

Miller K. (2010). Using a Computer-Based Risk Assessment Tool to Identify Risk for Chemotherapy-Induced Febrile Neutropenia. Clinical Journal of Oncology Nursing 14(1), 87-91.

O´Brien, C., Dempsey, O., & Kennedy, M. J. (2014). Febrile neutropenia risk assessment tool: improving clinical outcomes for oncology patients. Eur J Oncol Nurs, 18(2), 167-174.

Renner, P., Milazzo, S., Liu, J. P., Zwahlen, M., Birkmann, J., & Horneber, M. (2012). Primary prophylactic colony-stimulating factors for the prevention of chemotherapy-induced febrile neutropenia in breast cancer patients. Cochrane Database Syst Rev, 10, Cd007913.

Weycker, D., Li, X., Edelsberg, J., Barron, R., Kartashov, A., Xu, H., & Lyman, G. H. (2014). Risk and Consequences of Chemotherapy-Induced Febrile Neutropenia in Patients With Metastatic Solid Tumors. Journal of Oncology Practice, 11(1), 47-54.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

This measure was stratified for disparities by age, gender, and race/ethnicity. The results/scores are presented for these categories in Table 2.

Table 2. Rates by Age, Race/Ethnicity, and Gender for the Entire Sample

Category: Denominator / Numerator / Measure Rate

All Patients: 192 / 24 / 0.125

Age (years)
18 – 44: 27 / 6 / 0.222
45 – 64: 69 / 5 / 0.072
65 – 74: 63 / 8 / 0.127
75 – 84: 30 / 5 / 0.167
85+: 3 / 0 / 0

Race/Ethnicity
White, non-Hispanic: 111 / 17 / 0.153
Black, non-Hispanic: 16 / 0 / 0
Hispanic: 30 / 2 / 0.067
Other: 13 / 2 / 0.154
Unknown: 22 / 3 / 0.136

Gender
Female: 134 / 20 / 0.149
Male: 58 / 4 / 0.069

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
In a retrospective observational study of breast cancer patients based on 1994-2003 Medicare- Surveillance, Epidemiology, and End Results (SEER) data, Rajan et al. (2011) found that first-cycle CSF is used more frequently in white women than in nonwhite women. In the same study, the percentage of patients who received prophylactic CSF varied substantially by SEER geographic region. In another study of breast cancer patients based on 1998-2005 Medicare-SEER data, CSF use was significantly lower among black and Hispanic women than white women, and among women with a lower socioeconomic status score (based on census tract-level education, poverty level, and income data) (Hershman et al., 2012).

Citations
Hershman DL, Wilde ET, Wright JD, et al. (2012). Uptake and economic impact of first-cycle colony stimulating factor use during adjuvant treatment of breast cancer. J Clin Oncol. 30:806-812, 2012

Rajan SS, Lyman GH, Carpenter WR, et al. (2011). Chemotherapy characteristics are important predictors of primary prophylactic CSF administration in older patients with breast cancer. Breast Cancer Res Treat 127, 511-520.

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
In the United States, about 650,000 patients receive chemotherapy in outpatient oncology clinics each year (Halpern et al., 2008). FN occurs frequently after chemotherapy, with the incidence among patients with a solid tumor estimated to be 13.1-20.6% during the chemotherapy course and 3.1-7.4% in the first cycle (Weycker et al., 2015), and incidence among patients with lymphoma estimated at 36% (Bohlius et al., 2008). FN is often due to a serious infection, which often leads to hospital admission and treatment with empiric broad-spectrum antibiotics, and results in higher morbidity, mortality, and costs (Lyman, 2003). The risk of FN may also result in reductions and delays in chemotherapy doses, which may negatively affect longer-term clinical outcomes. Empirical estimates from studies of these effects are included in the Evidence Form (in a separate file as part of this submission).

The incremental cost of chemotherapy-related serious adverse effects (i.e., "expenditure in excess of those made on behalf of the matched non-chemotherapy recipients") in a population sample of women under 65 years of age with breast cancer has been estimated to be $1271 (in 2006 dollars) per person per year (Hassett et al., 2006). Extrapolating from this per-person cost estimate, Hassett et al. report the national incremental cost of chemotherapy-related serious adverse effects may be as high as $59.8 million per year (in 2015 dollars, inflated from 2006 to 2015 dollars using medical-cost inflation rates [Halfhill, 2016]) for this subgroup of patients. Based on data from the American Cancer Society, breast cancer cases are estimated to represent 16.8% of incident cases of solid malignant tumors that will occur during 2016 in the United States (American Cancer Society, 2016). Therefore, the total

incremental cost of chemotherapy-related serious adverse effects for all patients with solid malignant tumors may be as high as $354.9 million per year (in 2015 dollars). Another recent study of 26,628 hospitalizations for febrile neutropenia in patients with breast cancer estimated a mean length of stay of 5.7 days (95 % CI 5.5–5.9 days), and a mean cost of hospitalization of $37,087 (95 % CI $34,009–$40,165) (Pathak et al., 2015).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

American Cancer Society. Cancer Facts & Figures 2016. Table 1. Estimated Number of New Cancer Cases and Deaths by Sex, US, 2016. Atlanta: American Cancer Society; 2016. Available at http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf

Bohlius, J., Herbst, C., Reiser, M., Schwarzer, G., & Engert, A. (2008). Granulopoiesis-stimulating factors to prevent adverse effects in the treatment of malignant lymphoma. Cochrane Database Syst Rev(4), Cd003189.

Halfhill, T. R. (2016). Tom's Inflation Calculator. Retrieved from http://www.halfhill.com/inflation_js.html

Halpern, M. T., & Yabroff, K. R. (2008). Prevalence of outpatient cancer treatment in the United States: estimates from the Medical Panel Expenditures Survey (MEPS). Cancer Invest, 26(6), 647-651.

Hassett, M. J., O´Malley, A. J., Pakes, J. R., Newhouse, J. P., & Earle, C. C. (2006). Frequency and Cost of Chemotherapy-Related Serious Adverse Effects in a Population Sample of Women With Breast Cancer. Journal of the National Cancer Institute, 98(16), 1108-1117.

Lyman GH. Risk assessment in oncology clinical practice. From risk factors to risk models. Oncology (Williston Park). 2003;17:8-13.

Pathak R, Giri S, Aryal MR, Karmacharya P, Bhatt VR, Martin MG. (2015). Mortality, length of stay, and health care costs of febrile neutropenia-related hospitalizations among patients with breast cancer in the United States. Support Care Cancer 23:615–617

Weycker, D., Li, X., Edelsberg, J., Barron, R., Kartashov, A., Xu, H., & Lyman, G. H. (2014). Risk and Consequences of Chemotherapy-Induced Febrile Neutropenia in Patients With Metastatic Solid Tumors. Journal of Oncology Practice, 11(1), 47-54.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

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 subcriteria 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):*

Cancer, Cancer : Bladder, Cancer : Breast, Cancer : Colorectal, Cancer : Gynecologic, Cancer : Hematologic, Cancer : Liver, Cancer : Lung, Esophageal, Cancer : Pancreatic, Cancer : Prostate, Cancer : Skin

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Safety : Complications, Safety : Healthcare Associated Infections, Safety : Medication Safety

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.)*

A measure-specific Web page was not set up for this measure.

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_2930_Code_Sets_3-11-16_To_NQF.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients who had an FN risk assessment documented in the medical record prior to the first cycle of intravenous chemotherapy.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The time period is 12 months for the administrative data that are used to determine if the patient received chemotherapy in the 12 months before the first cycle of the current regimen of intravenous chemotherapy. The time period for identifying a patient with first-cycle chemotherapy in the medical record is any time during the measurement period (12 consecutive months). There is also a 30-day look-back in the medical record from the start of the first cycle of intravenous chemotherapy for identifying the FN risk assessment.

S.6. **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, 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.
The numerator is defined as patients with an FN risk assessment documented in the medical record within 30 days before the first cycle of intravenous chemotherapy. An FN risk assessment is defined as at least one of the following:

- Template in the record or evidence that an online tool was used to assess FN risk (e.g., a Febrile Neutropenia Risk Assessment Tool similar to that described in the study by O'Brien et al. [2014])
- FN risk of the planned regimen was noted as a percentage (e.g., >20%) OR noted qualitatively (e.g., "high FN risk")
- Patient factor(s) was noted as a contributor to elevated FN risk (e.g., "high FN risk due to advanced age and comorbidity")
- Justification for USE of CSF was documented (e.g., "high risk regimen, CSF support will be used;" "due to the presence of expanders and risk of infection, CSF will be used")
- Justification for NOT using CSF was documented (e.g., "due to patient's youth and excellent health, CSF support will not be used")

Citation
O´Brien, C., Dempsey, O., & Kennedy, M. J. (2014). Febrile neutropenia risk assessment tool: improving clinical outcomes for oncology patients. Eur J Oncol Nurs, 18(2), 167-174.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of patients 18 years of age or older with a solid malignant tumor or lymphoma receiving the first cycle of intravenous chemotherapy.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, 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)*

IDENTIFICATION OF PATIENTS WITH SOLID MALIGNANT TUMOR OR LYMPHOMA IN MEDICAL RECORDS

The time period is defined as any time during the measurement period (12 consecutive months). The denominator includes patients treated for a solid malignant tumor or lymphoma with first cycle of intravenous chemotherapy who meet the following conditions:

1. Patient was 18 years of age or older when first-cycle intravenous chemotherapy of the current regimen was initiated.
2. Patient's first-cycle intravenous chemotherapy was initiated any time during months 2 through 12 of the 12-month measurement period.
3. The treatment ordered was intravenous chemotherapy (see sheet labeled "IV Chemotherapy" in the attached Excel file for a list of CPT procedure codes for chemotherapy).
4. Patient was being treated for a solid malignant tumor or lymphoma (see sheets labeled "Denom Diagnoses ICD9," "Denom Diagnoses ICD10," and "Denom Diagnoses ICD9-ICD10" in the attached Excel file for a list of ICD-9-CM diagnosis codes, ICD-10 CM diagnosis codes, and a conversion table between ICD-9-CM and ICD-10-CM diagnosis codes, respectively).
5. Patient did not receive chemotherapy in the 12 months prior to the first cycle of chemotherapy.
6. Patients receiving experimental therapy or participating in clinical trials are not eligible because the trial protocol dictates CSF prophylaxis decisions.
7. Patients on weekly chemotherapy regimens are not eligible because the intervals between treatments are not long enough for CSF prophylaxis to have an effect.
8. Patients receiving concurrent radiation therapy (see sheet labeled "Radiation Therapy" in the attached Excel file for CPT codes) are not eligible because CSF prophylaxis is contraindicated for those patients due to the risk of irreversible stem cell damage. Patients who receive palliative local radiation for pain control are eligible.
9. Record of care was complete (e.g., provider notes prior to cycle #1 of chemotherapy are available).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

There are no denominator exclusions.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, 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.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*

Measure results may be stratified by:

- Age – Divided into five categories: 18-44, 45-64, 65-74, 75-84, and 85+ years
- Race/Ethnicity
- Gender
- Curative/adjuvant and palliative chemotherapy
- Periodicity of chemotherapy (2-, 3- and 4-week cycles)

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

No risk adjustment or risk stratification

If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
Denominator: Number of patients 18 years of age or older with a solid malignant tumor or lymphoma receiving the first cycle of intravenous chemotherapy.

Create Denominator:
1. Identify patients who received intravenous chemotherapy in an outpatient setting during the measurement year (see sheet labeled "IV Chemotherapy" in the attached Excel file for CPT procedure codes for chemotherapy).
2. Of patients identified in Step 1, keep only patients who were being treated for a solid malignant tumor or lymphoma (see sheets labeled "Denom Diagnoses ICD9," "Denom Diagnoses ICD10," and "Denom Diagnoses ICD9-ICD10" in the attached Excel file for a list of ICD-9-CM diagnosis codes, ICD-10 CM diagnosis codes, and a conversion table between ICD-9-CM and ICD-10-CM diagnosis codes, respectively).
3. Of patients identified in Step 2, keep patients who initiated the first cycle of intravenous chemotherapy between February 1 and December 31 of the measurement year.
4. Of patients identified in Step 3, keep those who were 18 years of age or older when first-cycle intravenous chemotherapy was initiated.
5. Of patients identified in Step 4, keep patients who did not receive chemotherapy in the 12 months prior to the initiation of the first cycle of chemotherapy. This is the denominator of the measure.

Numerator: Number of patients who had an FN risk assessment documented in the medical record prior to the first cycle of intravenous chemotherapy.

Create Numerator:
For patients in the denominator, identify those with an FN risk assessment documented in the medical record prior to the first cycle of intravenous chemotherapy. This is the numerator of the measure. Any of the following can be counted as evidence that a risk assessment for FN was performed:

- Template in the record or online tool was used to assess FN risk (e.g., a Febrile Neutropenia Risk Assessment Tool similar to that described in the study by O'Brien et al. [2014])
- FN risk of the planned regimen was noted as a percentage (e.g., >20%) OR noted qualitatively (e.g., "high FN risk")
- Patient factor(s) was noted as a contributor to elevated FN risk (e.g., "high FN risk due to advanced age and comorbidity")
- Justification for USE of CSF was documented (e.g., "high risk regimen, CSF support will be used;" "due to the presence of expanders and risk of infection, CSF will be used")
- Justification for NOT using CSF was documented (e.g., "due to patient's youth and excellent health, CSF support will not be used")

The measure is calculated as the numerator divided by the denominator.

Citation
O´Brien, C., Dempsey, O., & Kennedy, M. J. (2014). Febrile neutropenia risk assessment tool: improving clinical outcomes for oncology patients. Eur J Oncol Nurs, 18(2), 167-174.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data source for the measure is medical records in electronic or paper form. The instrument used to abstract the information from the medical record was developed for this measure and is attached as a file called "Measure Data Collection Tool" to the Appendix of this form. The field test data collection form is available from the developer upon request.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Other
If other: Outpatient chemotherapy clinic

S.28. **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.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_2930_Testing_Form_3-11-16_To_NQF.docx,NQF_2930_Testing_Form_4-8-16_To_NQF.docx

Measure Number (*if previously endorsed*)**:** 2930

Measure Title: Febrile Neutropenia Risk Assessment Prior to Chemotherapy

Date of Submission: 3/11/2016

Type of Measure: (see checked box below)

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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*).

Not applicable; an existing dataset was not used.

1.3. What are the dates of the data used in testing? April 2011 through February 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: community oncology clinic	☒ other: community oncology clinic

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*)

A total of five community oncology clinics were included in the testing. We aimed to identify clinics that represent real-world practice and selected them through telephone book and internet searches. All five clinics are independent practices and not affiliated with academic medical centers, but all of them participate in clinical trials. Their characteristics are shown in Table 1. The clinics are located in three states, two are in an urban area and three are suburban, and the clinical staff ranges in size from 3 to 18 oncologists.

Table 1. Characteristics of Community Oncology Clinics

Location	Type of Area	Practice Composition
Maryland	Urban	3 oncologists, 1 NP
New Jersey	Suburban	7 oncologists
California	Urban	6 oncologists, 2 NPs
California	Suburban	18 oncologists, 5 PAs
New Jersey	Suburban	3 oncologists

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*)

We provided the clinic with detailed inclusion criteria (age at least 18 years, solid tumor or lymphoma, initiating chemotherapy, and not participating in a clinical trial) and asked them to select 40 patients at random, for whom they

had complete records (i.e., records that covered the intake exam and evaluation and the time period up to the second chemotherapy cycle). We received 200 records, but eight patients were not eligible because of incomplete records (6), malignancy other than solid tumor or lymphoma (1), or concurrent radiation (1).

Thus, a total of 192 patients from 5 clinics were included in the testing and analysis. The number of patients by clinic ranged from 37 to 40. Table 2 presents information about the patients in the sample of medical records that was abstracted. Patients 45-64 and 65-74 years of age each comprised about a third of the sample. More than two-thirds of the sample was female. Almost 58 percent of the sample was identified as white, non-Hispanic. Half of the sample had private health insurance and another quarter had Medicare coverage only. The four most frequent cancers in the sample were breast, lymphoma, lung, and colon.

Table 2. Demographic and Clinical Characteristics of Patients

Characteristic	Number	Percent
Age		
18-44	27	14.1%
45-64	69	35.9%
65-74	63	32.8%
75-84	30	15.6%
85+	3	1.6%
Gender		
Male	58	30.2%
Female	134	69.8%
Race/Ethnicity		
White	111	57.8%
Hispanic	30	15.6%
African American	16	8.3%
Other/Unknown	35	18.2%
Insurance		
Medicare Only	50	26.0%
Medicaid Only	9	4.7%
Medicare and Medicaid	4	2.1%
Self-Pay	29	15.1%
Private Health Insurance	98	51.0%
Missing	2	1.0%
Cancer		
Breast	71	37.0%
Lymphoma	31	16.1%
Lung	27	14.1%
Colon	22	11.5%
Ovarian	10	5.2%
Endometrial	8	4.2%
Bladder	6	3.1%
Gastric	6	3.1%
Prostate	4	2.1%
Pancreatic	2	1.0%

Characteristic	Number	Percent
Sarcoma	1	0.5%
Brain	1	0.5%
Esophageal	1	0.5%
Unknown Primary	2	1.0%

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.

The data used for testing inter-rater reliability consisted of 50 medical records that were abstracted by two abstractors.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The patient-level sociodemographic variables that were analyzed were age, gender, and race/ethnicity.

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*)

Two abstractors were instructed to abstract the same randomly selected 50 medical records, ten records per clinic, for a 25 percent inter-rater reliability (IRR) sample. The IRR was estimated for scoring whether documentation of a febrile neutropenia risk assessment was in the medical record using the "kap" command in Stata Statistical Software, which calculates the kappa-statistic measure of interrater agreement when there are two unique raters and two or more ratings (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP).

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*)

The kappa statistic and percent agreement between coders for scoring whether documentation of a febrile neutropenia risk assessment was in the medical record is shown in Table 3 for each of the five clinics.

Table 3. Inter-rater Reliability for Scoring Febrile Neutropenia Risk Assessment in Medical Record, by Clinic

Site	Number of Medical Records in Inter-Rater Reliability Sample	Kappa Statistic (SE)	Agreement*
Clinic 1	10	1.0	100%
Clinic 2	10	0.783 (0.201)	90%
Clinic 3	10	1.0	100%
Clinic 4	10	1.0	100%
Clinic 5	10	1.0	100%
All sites	50	0.878 (0.120)	98%

*Percent of records for which the two coders agreed as to whether there was or was not a documented reference to FN risk in the record.

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?*)

For the records included in the inter-rater reliability sample, Kappa estimates ranged from 0.783 to 1.0 for the five clinics. For four of the clinics, the kappa estimates were 1.0, indicating there were no differences between coders in the scoring of whether there was or was documentation of an FN risk assessment in the record. For the other clinic, the kappa estimate was 0.78, indicating that there was little disagreement between the coders in the scoring of measure (Fleiss, 1981; Landis and Koch, 1977). The simple agreement between the codes was between 90 and 100 percent.

Citations

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley.
Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics 33 (1): 159–174.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

We identified ten experts in clinical oncology to rate the measure on its face validity and usability. The names and organizations of the clinical experts are listed in Table 4. The face validity was rated using a web-based questionnaire (developed using SurveyMonkey®). The clinical experts were asked to review the measure specifications and the evidence supporting the measure from the NQF forms, which we provided to them. After reviewing the background material, they were instructed to rate two statements about the measure by indicating their level of agreement on a 5-point scale (1=Strongly Disagree; 2=Disagree; 3=Neither Agree nor Disagree; 4=Agree; 5=Strongly Agree). The first statement was related to the face validity of the measure: “Performance scores resulting from the measure as defined

can be used to distinguish good from poor quality." The second statement was related to the usability of the measure: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)."

Table 4. Clinical Experts Who Rated Measure on Face Validity and Usability

Name	Organization
John Glaspy MD Los Angeles, CA	Chair, Founders Board and UCLA Leadership Professor of Medicine, Jonsson Comprehensive Cancer Center Estelle, Abe, and Marjorie Sanders Endowed Chair in Cancer Research Director, Jonsson Comprehensive Cancer Center Clinical Research Unit Director, Jonsson Comprehensive Cancer Center Women's Cancer Research Program Vice-Chair, Jonsson Comprehensive Cancer Center Scientific Protocol Review Committee
Hind Hamdan, MD Hagerstown, MD	Practicing Oncologist, Antietam Oncology, Hagerstown, MD
Thomas Lowe, MD Redondo Beach, CA	Vice-Chair of the Oncology Committee, Little Company of Mary Hospital Chairman, Breast Cancer Advisory Board for Little Company of Mary Hospital
Reshma Mahtani, DO Miami, FL	Assistant Professor of Clinical Medicine, Division of Hematology/Oncology Sylvester Comprehensive Cancer Center University of Miami Health System
Timothy Moore, MD Columbus, OH	Medical Director for the Palliative Care Service Grant Medical Center in Columbus
Vicki Morrison, MD Duluth, MN	Professor of Medicine in Hematology/Oncology and Infectious Disease University of Minnesota Medical School
Edgardo S. Santos, MD, FACP Boca Raton, FL	Associate Professor of Clinical Biomedical Sciences, Florida Atlantic University Medical Director of Cancer Research, Lynn Cancer Institute
Robert Smith, Jr., MD Columbia, SC	Clinical Associate Professor, University of South Carolina School of Medicine Practicing Oncologist, Kershaw Health and Lexington Medical Center
Charles Vogel , MD Miami, FL	Professor of Clinical Medicine, Division of Hematology/Oncology Sylvester Comprehensive Cancer Center University of Miami Health System
Andrew Zelenetz, MD New York, NY	Vice Chair of Medical Informatics Memorial Sloan Kettering Cancer Center

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

Ten clinical oncology experts completed the evaluation of the measure's face validity and usability. The results of the rating of face validity on a scale of 1 to 5 are presented in Table 5.

Table 5. Results of the Face Validity Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	1 (10.0%)
4 (Agree)	7 (70.0%)
3 (Neither Agree nor Disagree)	1 (10.0%)
2 (Disagree)	1 (10.0%)
1 (Strongly Disagree)	

Of the clinical oncology experts who rated the measure for face validity, 80.0% (8/10) strongly agreed or agreed with this statement: "Performance scores resulting from the measure as defined can be used to distinguish good from poor

quality". The mean rating was 3.8, and the median rating was 4. The results of the rating of usability on a scale of 1 to 5 are presented in Table 6.

Table 6. Results of the Usability Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	2 (20.0%)
4 (Agree)	5 (50.0%)
3 (Neither Agree nor Disagree)	1 (10.0%)
2 (Disagree)	1 (10.0%)
1 (Strongly Disagree)	1 (10.0%)

Of the clinical oncology experts who rated the measure for usability, 70.0% (7/10) strongly agreed or agreed with this statement: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)." The mean rating was 3.6, and the median rating was 4.

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

In summary, 80 percent of the clinical oncology experts who participated in the rating strongly agreed or agreed that the measure has face validity, and 70 percent strongly agreed or agreed that the measure exhibits usability. This indicates strong support for the validity and usability of the measure from the clinical oncology experts.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

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

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
Not applicable.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
Not applicable.

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

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)
Not applicable.

2b4.4a. What were the statistical results of the analyses used to select risk factors?
Not applicable.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)
Not applicable.

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable.

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable.

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

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in*

1b)

We calculated the rate of the FN risk assessment measure for each clinic and tested for statistically significant differences between the rates in the clinics using a test for the difference between two independent proportions. For those tests with an expected cell size of less than 5, we used a Fisher exact probability test. We classified differences between each pair of clinics as statistically significant based on two-tailed p-values.

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

The rate of having documentation of an FN risk assessment in the medical record ranged from 0 to 0.27 in the five clinics (Table 7). There are statistically significant differences in the measure rates between Clinic 4 and each of Clinics 1, 2, and 3 and between Clinic 5 and each of Clinics 1 and 2 (Table 8).

Table 7. Numerator, Denominator, and Measure Rate by Clinic

Clinic Number	Numerator	Denominator	Measure Rate
Clinic 1	6	37	0.162
Clinic 2	10	37	0.270
Clinic 3	7	39	0.179
Clinic 4	0	39	0.000
Clinic 5	1	40	0.025

Table 8. Statistical Significance of Comparisons Between Clinics Based on Two-Tailed Significance Test

	Clinic 1	Clinic 2	Clinic 3	Clinic 4
Clinic 1				
Clinic 2	NS*			
Clinic 3	NS*	NS*		
Clinic 4	P<0.05**	P<0.001**	P<0.05**	
Clinic 5	NS**	P<0.01**	P<0.05**	NS**

NS= Not significant at P<0.05
*Based on test for difference between two independent proportions.
**Based on a Fisher exact probability test.

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

We conclude that the differences are clinically meaningful, as two sites had a rate of zero or almost zero whereas others achieved almost a 0.3 rate. The results point to overall less-than-optimal performance, but also variation across sites. Based on the abstracted medical records in the field test, we were able to identify statistically significant differences in the documented FN risk assessment measure rate across community oncology clinics. An expanded field test or data from implementation in a state or national quality improvement program should provide power to identify statistically significant and clinically meaningful differences in performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)
Not applicable.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable.

2b6.3. What is your interpretation of the results in terms of 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*)
Not applicable.

_____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Not applicable. We did not identify missing data during the medical record abstraction.

2b7.2. 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. We did not identify missing data during the medical record abstraction.

2b7.3. 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. We did not identify missing data during the medical record abstraction.

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)

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)*

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.**

Information about FN risk assessment is often found in the patient's medical record and therefore may not be available from electronic sources.

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.**

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Testing demonstrated that the measure was feasible to specify and calculate using medical record data. Medical record data needed to implement the measure are available, accessible, and timely.

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)*.

There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

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.

Planned	Current Use (for current use provide URL)
Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. For each CURRENT use, checked above, provide:

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

Not applicable; the measure is being submitted for initial endorsement.

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

Not applicable; the measure is being submitted for initial endorsement.

4a.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.*)

Because the measure is being submitted to NQF for initial endorsement, we do not yet have specific plans to submit it for use in a specific federal, state or local program. However, this measure would be appropriate for use in a CMS reporting program for outpatient care provided to oncology patients, for example, under oncology bundled payment demonstrations. We will explore the possibility of submitting this measure through the Measures under Consideration process for the one of the CMS reporting programs. This would entail submitting information about the measure through JIRA, which is the CMS software system for collecting information on candidate measures for the list of "Measures under Consideration" for the annual pre-rulemaking process.

4b. 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.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Not applicable; information about progress on improvement is not required because this measure is being submitted for initial endorsement.

4b.2. **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.**
Not applicable; this measure is being submitted for initial endorsement.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
The measure has not been implemented in any reporting programs, and no unintended negative consequences were identified during testing.

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

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 completely harmonized?

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.)
Not applicable; there are no competing NQF-endorsed measures.

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:** NQF_2930_Measure_Data_Collection_Tool_4-8-16_To_NQF.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** RAND Corporation
Co.2 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622
Co.3 **Measure Developer if different from Measure Steward:** RAND Corporation
Co.4 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622

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.
A group of ten clinical oncology experts was used to rate the face validity and usability of the measure. Their names and affiliations are provided in the Testing Form.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Some proprietary codes are contained in the measure specifications for convenience of the user. Use of these codes may require permission from the code owner or agreement to a license.

ICD-10 codes are copyrighted © World Health Organization (WHO), Fourth Edition, 2010. CPT © 2010 American Medical Association. CPT is a registered trademark of the American Medical Association. All rights reserved.
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:** Liz Sloss at RAND made the revisions in the NQF submission form related to the field testing results on April 8, 2016, and re-submitted the measure.

Based on an email exchange between Amber Sterling at NQF and Liz Sloss at RAND (see below), this form will be updated after the March 11, 2016, submission deadline with results from the field test and other items related to testing.

_____
From: cancerem [mailto:cancerem@qualityforum.org]
Sent: Friday, February 26, 2016 2:42 PM
To: Sloss, Liz; cancerem
Cc: Mattke, Soeren; Roth, Carol; Qureshi, Nabeel
Subject: RE: Questions about measure submission for the Cancer 2015-2016 Call for Measures

Hi Liz,
I have spoken to our Senior Director on the project and she is comfortable with you all submitting the testing by Friday April 8th. Please try very hard to stick to this timeline as we will then internally review the measure before we submit it to our committee.

As it gets closer, we can discuss how you actually will submit the amended information. If you aren't able to add it directly to the submission, we can definitely do it for you.

Thanks,

Amber

_____

From: Sloss, Liz [mailto:sloss@rand.org]
Sent: Thursday, February 25, 2016 4:40 PM
To: cancerem
Cc: Mattke, Soeren; Roth, Carol; Qureshi, Nabeel
Subject: RE: Questions about measure submission for the Cancer 2015-2016 Call for Measures

Hi Amber,
We're planning to submit a new measure on "Febrile Neutropenia Risk Assessment Prior to Chemotherapy" to be considered for endorsement under the current Cancer 2015-2016 Call for Measures.
This measure is abstracted from electronic or paper medical records and we unfortunately encountered a few delays in obtaining enough records for the field test. So we're able to submit the Online Submission Form and the Evidence Form by the March 11 (6:00 PM ET) deadline, but will have to submit the Testing Form later. We would submit the Testing Form by Friday, April 8, at the latest. If the Testing Form is completed before that date, we will submit it as soon as it is completed. We have a few related questions:

- If the Measure Submission Form and the Evidence Form are submitted online by Friday, March 11, and the Testing Form is submitted by Friday, April 8, can the new measure be considered for endorsement under the Cancer 2015-2016 Call for Measures?
- How should we submit the testing form after the Friday, March 11 (6:00 PM ET) submission deadline?
  - Can we access the already submitted Measure Submission Form ourselves and upload the testing form as an attachment?
  - If not, can we email the testing form to you, so you can attach it manually to the already submitted Measure Submission Form?

I really appreciate your help in navigating the NQF Submission process.
Thanks,
Liz

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 #: 2936
Measure Title: Admissions and Emergency Department (ED) Visits for Patients Receiving Outpatient Chemotherapy
Measure Steward: Centers for Medicare & Medicaid Services (CMS)
Brief Description of Measure: Measure estimates hospital-level, risk-adjusted rates of inpatient admissions or ED visits for cancer patients >18 years of age for at least one of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of hospital outpatient chemotherapy treatment. The two rates are calculated and reported separately.
Developer Rationale: Chemotherapy treatment can have severe, predictable side effects, which, if inappropriately managed, can reduce patients' quality of life and increase healthcare utilization and costs. On average, cancer patients receiving chemotherapy have one hospital admission and two ED visits per year; approximately 40 percent of these admissions, and 50 percent of these ED visits stem from complications of chemotherapy, respectively [1]. This measure aims to assess the care provided to cancer patients and encourage quality improvement efforts to reduce the number of potentially avoidable inpatient admissions and ED visits among cancer patients receiving chemotherapy in a hospital outpatient setting. Improved hospital management of these potentially preventable symptoms—including anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—could reduce admissions and ED visits and increase patients' quality of care and quality of life.

Admissions and ED visits are costly to payers and reduce patients' quality of life. Measuring potentially avoidable admissions and ED visits for cancer patients receiving outpatient chemotherapy will provide hospitals with an incentive to improve the quality of care for these patients by taking steps to prevent and better manage side effects and complications from treatment. Hospitals that provide outpatient chemotherapy should implement appropriate care to minimize the need for acute hospital care for these adverse events. This measure would encourage hospitals to use guidelines from the American Society of Clinical Oncology, National Comprehensive Cancer Network, Oncology Nursing Society, Infectious Diseases Society of America, and other professional societies with evidence-based interventions to prevent and treat common side effects and complications of chemotherapy. This risk-standardized measure seeks to increase transparency in the quality of care patients receive and to provide information to help physicians and hospitals mitigate patients' need for acute care, which can be a burden on patients, and increase patients' quality of life.

This measure is envisioned to meet two National Quality Strategy priorities: (1) promoting effective communication and coordination of care, and (2) promoting the most effective prevention and treatment practices for the leading causes of mortality.

Citation

1. Vandervelde A, Miller H, Younts J. Impact on Medicare payments of shift in site of care for chemotherapy administration. Washington, DC: Berkeley Research Group; June 2014. http://www.communityoncology.org/UserFiles/BRG_340B_SiteofCare_ReportF_6-9-14.pdf. Accessed September 16, 2015.

Numerator Statement: This measure involves calculating two mutually exclusive outcomes: one or more inpatient admissions or one or more ED visits for any of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever,

nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of chemotherapy treatment among cancer patients receiving treatment in a hospital outpatient setting. These 10 conditions are potentially preventable through appropriately managed outpatient care. The qualifying diagnosis on the admission or ED visit claim must be (1) the principal diagnosis or (2) a secondary diagnosis accompanied by a principal diagnosis of cancer.
Denominator Statement: The measure cohort includes Medicare FFS patients aged 18 years and older as of the start of the performance period with a diagnosis of any cancer who received at least one hospital outpatient chemotherapy treatment at the reporting hospital during the performance period.
Denominator Exclusions: We established the following exclusion criteria after reviewing the literature, examining existing measures, reviewing feedback from a public comment period, and discussing alternatives with the Cancer Working Group and TEP members (see Section Ad.1. for description of group and membership). The goal was to be as inclusive as possible; we excluded only those patient groups for which hospital visits were not typically a quality signal or for which risk adjustment would not be adequate. The exclusions, based on clinical rationales, prevent unfair distortion of performance results.

1) Patients with a diagnosis of leukemia at any time during the performance period.

Rationale: Patients with leukemia are excluded due to the high toxicity of treatment and recurrence of disease so that admissions do not reflect poorly managed outpatient care for this population. Patients with leukemia have an expected admission rate due to relapse, so including leukemia patients in the cohort could be conceptualized as a planned admission, which does not align with the intent of the measure.

2) Patients who were not enrolled in Medicare FFS Parts A and B in the year prior to the first outpatient chemotherapy treatment during the performance period.

Rationale: We exclude these patients to ensure complete patient diagnosis data for the risk-adjustment model, which uses the year prior to the first chemotherapy treatment during the period to identify comorbidities.

3) Patients who do not have at least one outpatient chemotherapy treatment followed by continuous enrollment in Medicare FFS Parts A and B in the 30 days after the procedure.

Rationale: We exclude these patients to ensure full data availability for outcome assessment.

Measure Type: Outcome
Data Source: Administrative claims
Level of Analysis: Facility

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report
1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer provided a diagram of the link between the health outcome and the healthcare processes that influence it.
- The developer stated that the management of symptoms associated with outpatient chemotherapy is expected to reduce the risk of admissions and ED visits for side effects and complications such as nausea and vomiting, anemia, and neutropenic fever. Treatment plans and guidelines exist to support the management of these

conditions. Hospitals that provide outpatient chemotherapy should implement appropriate care to minimize the need for acute hospital care for these adverse events.

- The developer provided details on options to prevent and manage the side effects and symptoms of cancer and outpatient chemotherapy treatment to decrease the risk of admissions and ED visits.

Guidance from the Evidence Algorithm: Health outcome measure→The relationship between the outcome and at least one process is identified and supported by the stated rationale→Pass

Question for the Committee:

- *Is there at least one action the provider can take to achieve a change in the measure results?*

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

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

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

- The developer provided the following performance data in hospital-level variation from July 1, 2012 to June 30, 2013 using Medicare FFS claims:

	# hospitals	Range	Median	25th and 75th percentile
# hospitals	3,765	--	--	--
# patients	252,408	--	--	--
Risk-standardized **inpatient admission** rate	--	6.0% - 24.9%	10.2%	9.8% (25th) 10.8% (75th)
Risk-standardized **ED visit** rate	--	2.1% - 7.5%	4.1%	4.0% (25th) 4.4% (75th)

- The developer provided additional data addressing the gaps in outpatient chemotherapy care from the literature.

Disparities:

- The developer did not provide disparities data from the measure as specified but did examine associations between outcomes and SDS factors. The developer analyzed dual-eligibility, race, and AHRQ SES Composite Index to determine if these factors affected whether patients receiving hospital-based outpatient chemotherapy were more likely to have an inpatient admission and emergency department visit within 30 days than "non-low SDS" patients.
- On a patient-level, the developer's analysis found disparities based on the three variables examined. However, theses disparities were no longer significant when evaluated on the hospital-level. See SDS Testing for details.

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)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims

Specifications:

- This is a facility-level measure
- The numerator includes: one or more inpatient admissions OR one or more ED visits for any of the following 10 qualifying diagnoses - anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis
    - The time frame includes the 30 days following the date of each chemotherapy treatment (including the day of treatment) in an outpatient setting; inpatient admission or ED visit must occur within 30 days of chemotherapy treatment
    - 1 of the 10 qualifying diagnoses on inpatient admission or ED visit claims must be (1) the principal diagnosis or (2) secondary diagnosis accompanied by a principal diagnosis of cancer
    - Patients with both an inpatient admission and an ED visit during the performance period are counted towards the inpatient admission 'outcome' only. [*The developer states that the rates are calculated separately because the severity and cost of an inpatient admission is different from that of an ED visit, but both adverse events are important signals of quality and represent patient-important outcomes of care.]*
    - The inpatient admission or ED visit (outcome) is attributed to the hospital where the patient received chemotherapy treatment during the 30 days prior to inpatient admission or ED visit.
        - IF the patient received outpatient chemotherapy treatment from more than one hospital in the 30 days prior to inpatient admission or ED visit, then the admission/visit (outcome) is attributed to all the hospitals that provided treatment in those 30 days
- The denominator includes: Patients 18 years and older with a diagnosis of cancer who received at least one hospital outpatient chemotherapy treatment at the reporting hospital during the performance period
- Denominator exclusions include:
    - Patients with leukemia
    - Patients who were not enrolled in Medicare FFS Parts A & B in the year prior to the first outpatient chemotherapy treatment
    - Patients who do not have at least one outpatient chemotherapy treatment followed by continuous enrollment in Medicare FFS Parts A & B in the 30 days after the procedure
- A 'coding crosswalk' between ICD-9-CM codes and ICD-10-CM codes is included. Data dictionary available on Sharepoint site
- The measure is risk-adjusted
- A calculation algorithm is provided

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Do you agree with the 30 day time frame? Do you agree that the inpatient admission/ED visit should be attributed*

to ALL hospitals that provided outpatient chemotherapy treatment within the time frame?

- *Do you agree that the rates should be calculated separately? Should they be equally weighted?*

2a2. Reliability Testing Attachment

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.

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The dataset used included 2012-2013 Medicare data from 3,765 hospitals and 240,446 patients . A total of 942 hospitals with ≥ 60 patients in the cohort were included in the sample. [*Note: The developer states that only hospitals with at least 60 patients (30 patients in each of the split sample) were included in the sample.*]
- The developers used a split-sample methodology to test the measure score reliability; this is an appropriate method. Developers randomly assigned half of the patients in each hospital to two separate groups, calculated the performance measure score for each hospital in each of the two groups, and calculated the Pearson correlation between the performance rates; the higher the correlation, the higher the reliability of the measure.
  - Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Guide from Evans (1996) suggests for the absolute value of r: 0.00-0.19 as "very weak", 0.20-.039 "weak",0.40-0.59 "moderate", 0.60-0.79 "strong", and 0.80-1.0 "very strong". [Evans, J.D. (1996) Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove.]
- The developers also used the intraclass correlation coefficient (ICC) signal-to-noise method to determine the recommended minimum number of cases needed to maintain a reliability level of 0.4 or higher. The ICC reflects the percentage of variance in score results that is due to "true" or real variance between the hospitals.

Results of reliability testing :

- Measure score reliability results:
  - Inpatient admissions: **0.41** (95% (CI) = 0.37-0.45)
  - ED visits: **0.27** (95% (CI) = 0.22-0.33).
  - To achieve a reliability (ICC) of 0.4, a minimum of **25** patients are required to calculate the **inpatient admissions rate** and a minimum of **20** patients for the **ED visit rate** per performance period. [*The developer recommends a performance period long enough to accumulate a sufficient number of patients per hospital for improved reliability.*]
  - Pearson correlation between the performance rates not provided.

Guidance from the Reliability Algorithm: Precise specifications (Box 1)→empiric reliability testing (Box 2)→performance measure score (Box 4)→random split-half correlation and signal-to-noise methodologies used to calculate reliability rates for inpatient admissions and ED visits (Box 5)→

- moderate certainty/confidence that performance measure scores for **inpatient admissions** are reliable (Box 6b)→Moderate
- low certainty/confidence that performance measure scores for **ED visits** are reliable (Box 6c)→Low

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability – inpatient admissions: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for reliability – ED visits: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer states that face validity was assessed by external groups including their technical expert panel (TEP) of national experts and stakeholder organizations.
- The developer used the 2012-2013 Full Sample dataset to conduct a construct validation analysis of hospital attribution and the outcome definition.
    - This measure assumes that the hospitals administering outpatient chemotherapy are also responsible for managing the patient's clinical care and treatment-related complications. The developers assessed the extent to which patients received their outpatient chemotherapy at one hospital rather than across multiple hospitals.
    - The developer calculated the frequency of each qualifying diagnosis (anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, and sepsis) and compared it with the expectations of clinical and coding experts.
    - The developer also assessed potential differences in coding practices across hospitals by testing the frequency of the qualifying diagnoses when they are:
        - listed as principal diagnosis only;
        - listed as principal or secondary diagnosis;
        - listed as principal diagnosis OR as secondary diagnosis with a principal diagnosis of cancer.

Validity testing results:

- Face validity results: Stakeholders, including all members of the TEP, affirmed that the measure, as specified, can be used to distinguish between better and worse quality hospitals.
- Only **5%** of cancer patients in the cohort received outpatient chemotherapy from more than one hospital. The results demonstrated that most patients receive their treatment at one hospital and attributing the management of outpatient care to the reporting hospital is appropriate, according to the developer.
- Testing results demonstrated the most common reasons for inpatient admissions: pneumonia (21%), pain (19%), and anemia (12%). The most common reasons for ED visits included: pain (56%), fever (9%), and dehydration (8%).

- Less than one percent of patients were admitted or seen in the ED with a principal diagnosis of neutropenia or sepsis. The TEP and billing experts confirmed that the primary reason for inpatient admission or ED visit for these patients would likely be infection not neutropenia. To capture these patients pneumonia and sepsis were included in the qualifying diagnoses.

- The developers noted that when one of the ten qualifying diagnoses were coded as a secondary diagnosis with a principal diagnosis of cancer, the inpatient admission/ED visit observed rates were slightly higher than when the diagnoses were limited to the principal diagnosis only: 10.3% vs. 7.2% for inpatient stays and 4.2% vs. 3.9% for ED visits.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Are the potential differences in coding practices adequately addressed?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- A total of 80,070 unique patients, or about 25% of the eligible cohort (320,516 patients), were excluded due to:
  - Diagnosis of leukemia: 25,714 (8%) patients
  - Patients who were not enrolled in Medicare FFS Parts A and B 12 months prior to first chemotherapy treatment: 55,926 (17%) patients
  - Patients who did not have at least one chemotherapy treatment with enrollment in Medicare FFS Parts A and B 30-days after the procedure: 18,193 (6%) patients
- The developers *initially* excluded patients 18-64. An analysis of this cohort found that there was not a strong statistical or clinical reason to exclude younger patients from the measure cohort – all adult patients 18 years and older remain in the eligible cohort.
- The developer calculated observed rates prior to exclusions and found higher inpatient admission rates (14.1% vs. 10.3%) as expected, primarily due to leukemia patients. *[Leukemia patients have an expected admission rate due to relapse].*
- The table below indicates that there is modest variation in the number of cases excluded within hospitals. The developers state that due to this variation, failure to exclude the patients below may distort the measure score and unfairly disadvantage certain hospitals.

Table 1. Distribution of Percentage of Patients Excluded across Hospitals (N=3,765 hospitals)

Exclusion	25th percentile	50th percentile	75th percentile
Diagnosis of leukemia	0	5.1	9.5
12 month prior enrollment	0	12.5	20.7
30-day continued enrollment	0	1.5	7.0

- The developer states that the measure captured 75.0% of qualifying patients after all exclusions were applied.
- The developer also states that the exclusions are clinically relevant or required for data completeness to calculate risk-adjustment and identification of admission or ED visits.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure? Do you agree that leukemia patients should be excluded from the measure?*
- *Do you agree that the exclusions are of sufficient frequency and variation across hospitals to be needed?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

Description of the model

- The measure is risk-adjusted using a hierarchical logistic regression model for each outcome – inpatient admissions and ED visits. The same variable selection process was used for both models.
- The developers limited the initial selection of candidate variables for inclusion in the models to variables with a strong clinical rationale as identified in the literature and clinical expertise.
    - Candidate variables included demographic variables, comorbidities, indicators of disease severity (cancer type), exposure (# of chemotherapy treatments), and interactions (age-cancer type interaction).
- To select the final variables in each model, the developers removed the least significant variable one at a time until only statistically significant ($p<0.05$) variables remained in each model. Interactions between age and cancer type were retained in the models if they met a higher threshold for statistical significance ($p<0.01$) to increase the likelihood of being true interactions. The final risk-adjustment model for inpatient admissions includes 20 variables; the ED visits includes 15 variables.

Performance of the models

Discrimination statistics:

- The c-statistic, predictive ability, and over-fitting indices were calculated separately for each model in the development sample then compared with its performance in the validation sample; this was done separately for each measure.
- The c-statistic reflects how accurately a statistical model is able to distinguish between a patient with an outcome and a patient without an outcome. C-statistic values can range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of patients with and without the outcome of interest and a value of 1.0 indicates that the model perfectly identifies those with and without the outcome of interest . Generally, a c-statistic of at least 0.70 is considered acceptable.
- A wide range between the lowest decile and highest decile indicates the predictive ability to distinguish between high-risk and low-risk patients.
- The discrimination statistical results for the models in both measures are:
    - **Inpatient admission outcome model**
        - Development sample results:
            - c-statistic: 0.73
            - predictive ability: 2.09-27.70%
        - Validation sample results:
            - c-statistic: 0.72
            - predictive ability: 2.16-27.98%
    - **ED visit outcome model**
        - Development sample results:
            - c-statistic: 0.63
            - predictive ability: 1.91-8.33%
        - Validation sample results:
            - c-statistic: 0.64
            - predictive ability: 1.93-8.22%

Calibration statistics:

- The developers assessed model calibration by calculating over-fitting indices for each of the models in both measures. Over-fitting refers to the phenomenon in which a model describes the relationship between predictive variables and outcome(s) in one group of patients but fails to provide valid predictions in another distinct group of patients. Calibration values far from 0 and 1 provide evidence of over-fitting.
- The calibration statistical results for the models in both measures are:
    - **Inpatient admission outcome model**
        - Development sample results:
            - Calibration: (0,1)
        - Validation sample results:
            - Calibration: (0.01, 1.00)
    - **ED visit outcome model**
        - Development sample results:

        - Calibration: (0,1)
    - Validation sample results:
        - Calibration: (-0.04, 0.99)
- The developers noted that the risk-decile plots for both measures demonstrated that the models showed very similar results.

Conceptual basis and empirical support for potential inclusion of SDS factors in risk-adjustment approach

- The developer performed a literature search to help inform their conceptualization of the pathways by which SDS factors affect admissions and ED visits for patients receiving chemotherapy treatment in the hospital outpatient setting. Studies indicated that individuals that identify as a racial minority, with low socioeconomic status (SES), with charity care or self-pay insurance, are women, or are unmarried were more likely to experience a gap in cancer care in the outpatient chemotherapy setting than their counterparts.
- The developer identified several potential conceptual pathways to consider:
    - Relationship of SDS with health
    - Access to care
    - Differential care across hospitals
- Based on their interpretation of the literature and analysis of the above pathways, the developers identified 3 SDS variables for potential inclusion in the risk-adjustment model:
    - Race (black, other) **[*Per NQF's Expert Panel on Risk Adjustment for Sociodemographic Factors, race and ethnicity are not and should not be used as proxies for SES; rather, their effects are confounded by SES (p. 42).*]**
    - Medicaid dual eligible status
    - AHRQ SES Composite Index
- Prevalence of these 3 SDS factors varied substantially (0-100%) across hospitals and were associated with the measured outcome.
- On the patient-level, patients with "low SDS" (Medicaid dual-eligibility, race as black, and AHRQ SES Composite Index) receiving hospital-based outpatient chemotherapy were more likely to have an inpatient admission and emergency department (ED) visit within 30 days than "non-low SDS" patients. See Section 1b.4 for details.
- On the hospital-level, there was no clear relationship between the median risk-standardized rates and hospitals' case mix by these three SDS factors. In addition, the analysis demonstrated that hospitals with a greater percentage of "low SDS" patients had similar rates of inpatient admissions and ED visits within 30 days of hospital-based outpatient chemotherapy. For example, hospitals in the lowest quartile of proportion of black patients had a median risk-adjusted admission rate of 10.2, the second quartile had a rate of 10.6, third quartile had a median rate of 10.1, and the top quartile of hospitals with proportion of black patients had a rate of 10.2.
- The risk-adjustment models demonstrated similar performance with and without including SDS variables in the methodology:
    - The inpatient admission measure c-statistics were 0.725 for the model that **did not** adjust for SDS variables and 0.728 for the model that adjusted for SDS variables. For the ED visit measure, the c-statistics were 0.636 **without** adjusting for SDS and 0.644 when adjusting for SDS.
    - The developer states that a very high agreement of hospital rankings between risk-adjustment models which included SDS variables and those that did not (Spearman rank correlation = 0.988 for the inpatient admission model and 0.984 for the ED visit model), suggested that accounting for SDS factors did not have a major impact on hospital rankings.
- Based on these results, the developer decided **NOT** to include any of the 3 SDS factors analyzed in the final risk-adjustment model.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment models adequately described for the measure to be implemented?*
- *Do you agree with the developer's decision, based on their analyses, to not include SDS factors in their risk-adjustment models?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developers provided hospital-specific risk-adjusted rates of potentially preventable inpatient admissions or ED visits for cancer patients aged 18 years or older receiving chemotherapy treatment in the hospital outpatient setting using a hierarchical logistic regression model.

Table 2. Outcome rates, among hospitals with any case size

Result	Mean	Stnd Dev	Min	25th Pctl	Median	75th Pctl	Max
Observed Admission Rate	8.3	0.11	0.0	0.0	6.5	12.1	100.0
Risk-Adjusted Admission Rate	10.4	1.28	6.0	9.8	10.2	10.8	24.9
Observed ED Visit Rate	4.3	0.09	0.0	0.0	1.4	5.3	100.0
Risk-Adjusted ED Visit Rate	4.2	0.53	2.1	4.0	4.1	4.4	7.5

Source: Claims from Medicare Parts A and B from July 1, 2012 through June 30, 2013.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

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

2b7. Missing Data

- N/A – the developers did not perform an empirical analysis of missing data or provide a rationale for a selected approach for handling missing data.

Guidance from Validity Algorithm **for inpatient admission rates**: Precise specifications (Box 1)→ potential threats to validity assessed /risk model performance acceptable (Box 2)→ face validity assessed (Box 4)→performance measure as specified can be used to distinguish quality (Box 5) →Moderate

Preliminary rating for inpatient admission rates validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from Validity Algorithm for **ED visit rates**: Precise specifications (Box 1)→ potential threats to validity assessed /risk model performance below what is considered acceptable (Box 2)→ face validity assessed (Box 4)→performance measure as specified can be used to distinguish quality (Box 5) →Low

Preliminary rating for ED visit rates validity: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

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 electronic claims and generated or collected by and used by healthcare personnel during the provision of care. The data are coded by someone other than person obtaining original information.
- Measure development and testing showed that the measure cohort can be defined and outcomes reported using routinely collected Medicare claims data. This measure is not in operational use.
- There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

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

Criterion 4: Usability and Use

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details:

- This is a recently developed measure that is not currently publicly reported or used in an accountability program. The measure may ultimately be used in one or more CMS programs such as Hospital Outpatient Quality Reporting Program and PPS-Exempt Cancer Hospital Quality Reporting Program.
- For the two CMS programs listed, the developer did not include the purpose, intended audience, and timeline for implementing the measure within the specified timeframes as required for measures not currently publicly reported or used in at least one accountability application.

Improvement results:

- Since this measure is not yet in use, there are no performance results to assess improvement. However, the developer expects there to be improvement in measure scores over time since publicly reported measure scores can reduce adverse patient outcomes associated with poorly managed outpatient care by capturing and making more visible to providers and patients all potentially preventable hospital visits following chemotherapy treatment in the hospital outpatient setting.

Feedback :

- In its 2015-2016 pre-rulemaking deliberations, the Measure Applications Partnership (MAP) conditionally supported this measure for the Prospective Payment System (PPS)-Exempt Cancer Hospital Quality Reporting (PCHQR) program. MAP advised that the measure undergo review and endorsement by NQF, with a special consideration from the Standing Committee of the exclusions and risk-adjustment methods.

Potential harms:

- The developer did not identify any unintended consequences during measure development or model testing. However, during the NQF Measure Applications Partnership (MAP) review of this measure in December 2015, there were concerns about a possible unintended consequence related to treatment decisions and underuse of appropriate care. The concern was that the measure might indirectly discourage more aggressive treatment plans that would have had clinical benefits. However, the purpose of the measure is to open lines of communication between the patient and provider on risks and preventative actions that can be taken for each type of treatment, and set the expectations for the patient so they can make more informed decisions on healthcare utilization as well. Furthermore, the measure is risk adjusted to help account for the variation in patient mix and aggressiveness of treatment. Lastly, the measure rate is not intended to be zero and CMS recognizes that not all admissions and ED visits are avoidable. Improving patient/provider communication and appropriately adjusting the model mitigates the risk of the unintended consequences.

***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

Criterion 5: Related and Competing Measures

Related or competing measures:

- 0383 : Oncology: Plan of Care for Pain – Medical Oncology and Radiation Oncology (paired with 0384)
- 0384 : Oncology: Medical and Radiation - Pain Intensity Quantified
- 1628 : Patients with Advanced Cancer Screened for Pain at Outpatient Visits

- Cancer – fatigue/anemia: percentage of patients seen for an initial visit or any visit while undergoing chemotherapy at a cancer-related outpatient site for whom there was an assessment of the presence or absence of fatigue (RAND Corporation) – this measure is not NQF endorsed.

Harmonization:

- All four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) focus on cancer patients receiving outpatient chemotherapy; however, there are some key differences in measure scope and measure type.
  - Measure scope: Each of the four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) narrowly focuses on pain management and/or fatigue/anemia. Measure 2936 does not target a specific symptom, but rather assesses the overall management of 10 important symptoms and complications that was identified as being more frequently cited in literature as reasons for ED visits and inpatient admissions following outpatient chemotherapy.
  - Measure type: The four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) are all process measures encouraging the use of screening and care plans to improve care. Measure 2936 is an outcome measure not encouraging or measuring specific processes to detect and treat these conditions, but rather assessing the outcomes of the care being provided. The four process measures, which are not risk-adjusted, support the intent of the measure by reinforcing that those providing outpatient care should screen for and manage symptoms such as pain.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Admissions and Emergency Department (ED) Visits for Patients Receiving Outpatient Chemotherapy

Date of Submission: 3/11/2016

Instructions

- *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.*
- Respond to all questions as instructed with answers immediately following the question. 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.
- Maximum of 10 pages (*incudes 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.

Note: The information provided in this form is intended to aid the Steering 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:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- 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.

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 U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

☒ Health outcome: One or more inpatient admissions or one or more emergency department (ED) visits for one of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of chemotherapy treatment among cancer patients receiving treatment in a hospital outpatient setting.

☐ Patient-reported outcome (PRO): Click here to name the PRO
PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

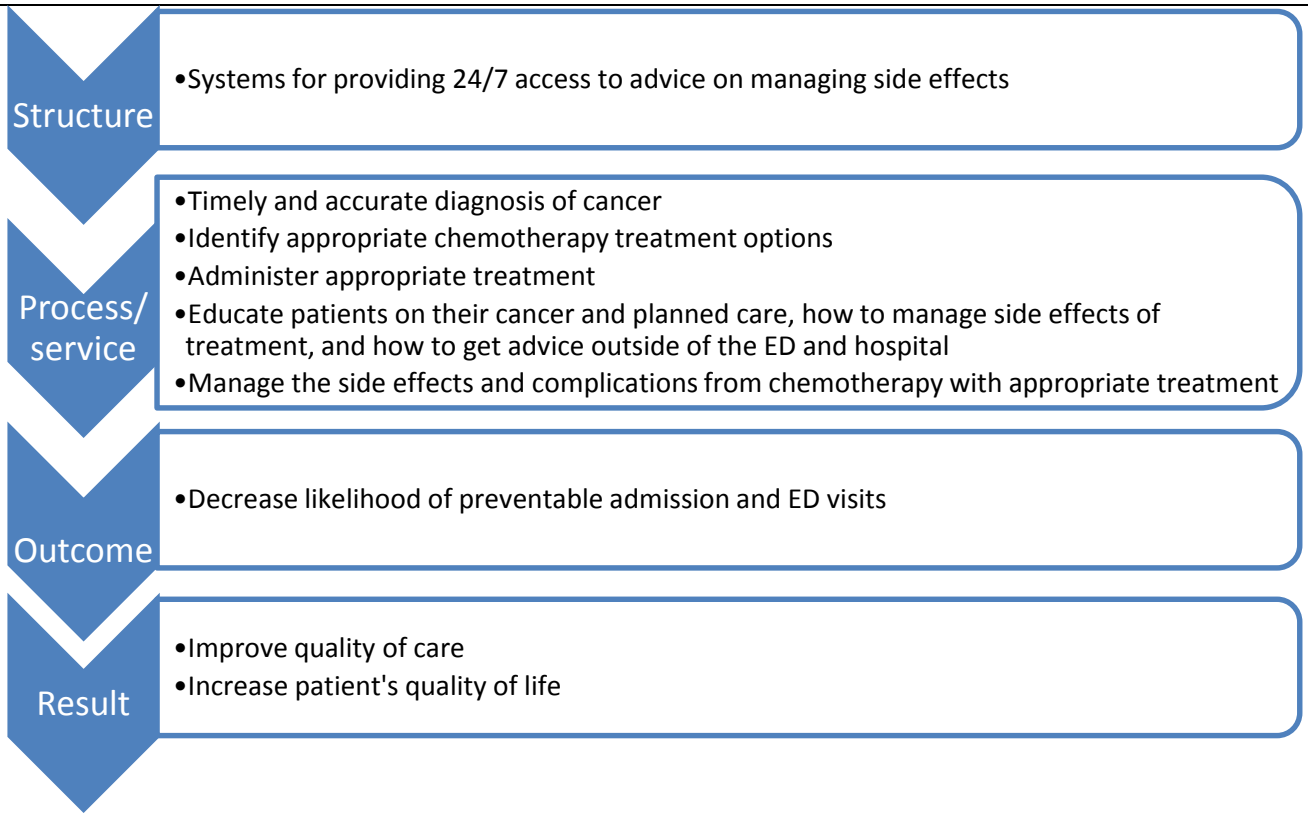
☐ Other: Click here to name what is being measured

_____

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

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

Unmet patient needs resulting in admissions and emergency department (ED) visits related to chemotherapy treatment pose a heavy financial burden and affect patients' quality of life. This measure will assess the percentage of cancer patients aged 18 years or older receiving hospital outpatient chemotherapy who have an admission or emergency department (ED) visit for anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis. This measure promotes proactive disease and symptom management for patients receiving hospital outpatient chemotherapy and is intended to encourage management of potentially preventable side effects and complications that could lead to hospital admissions or ED visits. Properly managing these side effects in the hospital outpatient setting and decreasing admissions and ED visits will reduce health care spending, improve quality of care, and increase the patient's quality of life. Below is a diagram of the path between the health outcome and the healthcare structures, processes, and results.



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

Chemotherapy treatment can have severe, predictable side effects, and hospital admissions and ED visits among patients receiving treatment in a hospital outpatient setting are often caused by manageable side effects and complications. Admissions and ED visits for eligible diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—may be due to patients receiving treatment in a hospital outpatient setting having unmet needs and gaps in care, which, if addressed, could reduce admissions and ED visits and increase patients' quality of life [7] [10] [11]. This gap in care may be due to reasons including (1) delayed onset of side effects that patients must manage at home, (2) patients' assuming that little can be done and not seeking assistance, and (3) limited access to and communication with providers who can tailor care to the individual [11].

Treatment plans and guidelines exist to support the management of these conditions. Hospitals that provide outpatient chemotherapy should implement appropriate care to minimize the need for acute hospital care for these adverse events. Guidelines from the American Society of Clinical Oncology, National Comprehensive Cancer Network, Oncology Nursing Society, Infectious Diseases Society of America, and other professional societies recommend evidence-based interventions to improve the quality of disease and symptom management. Management of symptoms associated with outpatient chemotherapy is expected to reduce the risk of admissions and ED visits for side effects and complications such as nausea and vomiting, anemia, and neutropenic fever. Below we provide more detail on options to prevent and manage the side effects and symptoms of cancer and outpatient chemotherapy treatment to decrease the risk of admissions and ED visits.

Anemia: There are many therapeutic agents available to treat anemia as well as clinical guidelines on how to prevent and manage anemia in patients receiving chemotherapy treatment [6] [4].

Dehydration: Dehydration can be prevented by educating patients on the importance of fluid intake and monitoring patients that have reduced oral intake or appetite loss. Health care professionals should also closely monitor patients at risk for chemotherapy-induced diarrhea and vomiting for signs of dehydration [15].

Diarrhea: Providers can often treat chemotherapy-induced diarrhea on an outpatient basis, and effective treatment of diarrhea can prevent dehydration [15]. Existing evidence supports management of diarrhea, although evidence about prevention continues to evolve [16].

Nausea/emesis: Chemotherapy-induced nausea and vomiting can be prevented and effectively managed in the outpatient setting [17]. Studies and reviews have shown the effectiveness of specific drugs for prevention and management of nausea and vomiting resulting from particular chemotherapy regimens and their effects on quality of life [1] [9] [13] [14].

Neutropenic fever: A systematic review and meta-analysis of randomized controlled trials concluded that prophylactic granulocyte colony-stimulating factors significantly reduce neutropenic fever [8].

Pain: A number of pharmacological treatments for pain exist, including opioids. However, many patients receive inadequate analgesia [5] [18]. Optimal pain control can be achieved through combining pharmacological and non-pharmacological approaches, in addition to assessing and reassessing patients' pain [2].

Pneumonia/Sepsis: The relationship between neutrophil count and the risk of infection is well established and studies have shown that risk factors can be identified and appropriate prophylactic measures, such as use of colony-stimulating factor, implemented to prevent neutropenia and associated complications [3]. Because of this relationship and the need for lab results to confirm neutropenia, neutropenia is often captured on the claim as the related infection, such as pneumonia and sepsis. The measure includes pneumonia and sepsis as outcomes to capture the same population [3] [8].

Citations

1. Billio, A., E. Morello, and M.J. Clarke. “Serotonin Receptor Antagonists for Highly Emetogenic Chemotherapy in Adults.” Cochrane Database of Systematic Reviews. Available at http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD006272.pub2/abstract. Accessed on September 24, 2012.

2. Chapman, S. “Assessment and Management of Patients with Cancer Pain.” *Cancer Nursing Practice*, vol. 10, no. 10, 2011, pp. 28–36

3. Crawford, J.C., D.C. Dale, and G.H. Lyman. “Chemotherapy-Induced Neutropenia.” *Cancer*, vol. 15, 2004, pp. 228–237.

4. Crowley, K., and K. Augustin. “Chemotherapy-Induced Anemia.” *US Pharmacist*, vol. 28, no. 04, 2003.

5. Fisch, M.J., J.W. Lee, M. Weiss, L.I. Wagner, V.T. Chang, D. Cella, J.B. Manola, L.M. Minasian, W. McCaskill-Stevens, T.R. Mendoza, and C.S. Cleeland. “Prospective, Observational Study of Pain and Analgesic Prescribing in Medical Oncology Outpatients with Breast, Colorectal, Lung, or Prostate Cancer.” *Journal of Clinical Oncology*, vol. 30, no. 16, 2012, pp. 1980–1988.

6. Groopman, J.E., and L.M. Itri. “Chemotherapy-Induced Anemia in Adults: Incidence and Treatment.” *Journal of the National Cancer Institute*, vol. 91, 2000, pp. 1616–1634.

7. Hassett, M.J., J. O’Malley, J.R. Pakes, J.P. Newhouse, and C.C. Earle. “Frequency and Cost of Chemotherapy-Related Serious Adverse Effects in a Population Sample of Women with Breast Cancer.” *Journal of the National Cancer Institute*, vol. 98, no. 16, 2006, pp. 1108–1117.

8. Kuderer, N.M., D.C. Dale, J. Crawford, and G.H. Lyman. “Impact of Primary Prophylaxis with Granulocyte Colony-Stimulating Factor on Febrile Neutropenia and Mortality in Adult Cancer Patients Receiving Chemotherapy: A Systematic Review.” *Journal of Clinical Oncology*, vol. 25, 2007, pp. 3158–3167.

9. Lohr, L. “Chemotherapy-Induced Nausea and Vomiting.” *Cancer Journal*, vol. 14, no. 2, 2008, pp. 85–93.

10. Mayer, D.K., D. Travers, A. Wyss, A. Leak, A. Waller. “Why Do Patients with Cancer Visit Emergency Departments? Results of a 2008 Population Study in North Carolina.” *Journal of Clinical Oncology*, vol. 26, no. 19, 2011, pp. 2683–2688.

11. McKenzie, H., L. Hayes, K. White, K. Cox, J. Fethney, M. Boughton, and J. Dunn. “Chemotherapy Outpatients’ Unplanned Presentations to Hospital: A Retrospective Study.” *Support Care Cancer*, vol. 19, 2011, pp. 963–969.

12. National Comprehensive Cancer Network. “Prevention and Treatment of Cancer-Related Infections.” NCCN Clinical Practice Guidelines in Oncology Version 1.2013. 2013. Available at http://www.nccn.org/professionals/physician_gls/f_guidelines.asp. Accessed Sept. 26, 2014.

13. Navari, R.M. “Prevention of Emesis from Multiple-Day and High-Dose Chemotherapy Regimens.” *Journal of the National Comprehensive Cancer Network*, vol. 5, no. 1, January 2007, pp. 51–59.

14. Osoba, D., B. Zee, D. Warr, J. Latreilee, L. Kaizer, and J. Pater. “Effect of Postchemotherapy Nausea and Vomiting on Health-Related Quality of Life.” *Support Care Cancer*, vol. 5, 1997, pp. 307–313.

15. Richardson, G., and R. Dobish. “Chemotherapy Induced Diarrhea.” *Journal of Oncology Pharmacy Practice*, vol. 13, no.4, 2007, pp. 181–98.

16. Stein, A., W. Voigt, and K. Jordan. "Chemotherapy-Induced Diarrhea: Pathophysiology, Frequency, and Guideline-Based Management." *Therapeutic Advances Medical Oncology*, vol. 2, 2010, pp. 51–63.

17. Trigg M,E., and G.M. Higa. "Chemotherapy-Induced Nausea and Vomiting: Antiemetic Trials that Impacted Clinical Practice." *Journal of Oncology Pharmacy Practice*, vol. 16, no. 4, December 2010, pp. 233–244.

18. Wu, H.S., T. Natavio, J.E. Davis, and H.N. Yarandi. "Pain in Outpatients Treated for Breast Cancer: Prevalence, Pharmacological Treatment, and Impact on Quality of Life." *Cancer Nursing*, vol. 36, no. 3, 2013, pp. 229–235. Available at http://www.cancernursingonline.com/. Accessed September 24, 2012.

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

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

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

Not applicable. This is an outcome measure.

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

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

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

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

☐ Other – ***complete section 1a.8***

Not applicable. This is an outcome measure.

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

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

☐ Yes → ***complete section 1a.7***

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

Not applicable. This is an outcome measure.

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

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

Complete section 1a.7

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

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

Complete section 1a.7

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

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

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

Not applicable. This is an outcome measure.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not applicable. This is an outcome measure.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

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

Not applicable. This is an outcome measure.

_____

1a.8 OTHER SOURCE OF EVIDENCE

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

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

Not applicable. This is an outcome measure.

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

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

ChemoMeasure_NQF_evidence_attachment.docx

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., the benefits or improvements in quality envisioned by use of this measure)*
Chemotherapy treatment can have severe, predictable side effects, which, if inappropriately managed, can reduce patients' quality of life and increase healthcare utilization and costs. On average, cancer patients receiving chemotherapy have one hospital admission and two ED visits per year; approximately 40 percent of these admissions, and 50 percent of these ED visits stem from complications of chemotherapy, respectively [1]. This measure aims to assess the care provided to cancer patients and encourage quality improvement efforts to reduce the number of potentially avoidable inpatient admissions and ED visits among cancer patients receiving chemotherapy in a hospital outpatient setting. Improved hospital management of these potentially preventable symptoms—including anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—could reduce admissions and ED visits and increase patients' quality of care and quality of life.

Admissions and ED visits are costly to payers and reduce patients' quality of life. Measuring potentially avoidable admissions and ED visits for cancer patients receiving outpatient chemotherapy will provide hospitals with an incentive to improve the quality of care for these patients by taking steps to prevent and better manage side effects and complications from treatment. Hospitals that provide outpatient chemotherapy should implement appropriate care to minimize the need for acute hospital care for these adverse events. This measure would encourage hospitals to use guidelines from the American Society of Clinical Oncology, National Comprehensive Cancer Network, Oncology Nursing Society, Infectious Diseases Society of America, and other professional societies with evidence-based interventions to prevent and treat common side effects and complications of chemotherapy. This risk-standardized measure seeks to increase transparency in the quality of care patients receive and to provide information to help physicians and hospitals mitigate patients' need for acute care, which can be a burden on patients, and increase patients' quality of life.

This measure is envisioned to meet two National Quality Strategy priorities: (1) promoting effective communication and coordination of care, and (2) promoting the most effective prevention and treatment practices for the leading causes of mortality.

Citation
1. Vandervelde A, Miller H, Younts J. Impact on Medicare payments of shift in site of care for chemotherapy administration. Washington, DC: Berkeley Research Group; June 2014. http://www.communityoncology.org/UserFiles/BRG_340B_SiteofCare_ReportF_6-9-14.pdf. Accessed September 16, 2015.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We assessed hospital-level variation in performance score using a Medicare FFS dataset as described below (please see Measure Testing Form Section 1.2 and 1.7 for full description of the datasets used).

We estimated the measure score for hospitals using Medicare FFS claims with a performance period of July 1, 2012 to June 30, 2013. The total number of hospitals with at least one attributed patient was 3,765. The total number of patients meeting inclusion and exclusion criteria across these hospitals was 252,408. The rate of risk-standardized inpatient admission rate ranged from 6.0 to 24.9 percent (median 10.2, 25th and 75th percentiles are 9.8 and 10.8, respectively). The rate of risk-standardized ED visit rate ranged from 2.1 to 7.5 percent (median 4.1, 25th and 75th percentiles are 4.0 and 4.4, respectively).

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.
There is a need to address the gaps in outpatient chemotherapy care because a significant number of cancer patients experience chemotherapy-related inpatient admissions and ED visits each year. Cancer patients average two ED visits per year, about half of which are chemotherapy related [4]. Furthermore, a study conducted in a single large metropolitan hospital in Australia showed that approximately 45 percent of chemotherapy outpatients had at least one ED visit within six months of chemotherapy treatment [5]. Similarly, a study conducted in a single hospital in Italy using data from 2006–2008, found that of 1,431 patients with hospital visits, 43 percent had received chemotherapy in the past three months [1]. With nearly half of cancer patients' ED visits related to chemotherapy treatment, measuring admissions and ED visits for patients receiving outpatient chemotherapy provides an opportunity for reporting hospitals to take steps to prevent and improve management of side effects and complications from treatment.

Among cancer outpatients, the frequently reported side effects of chemotherapy align with the reasons for hospital admissions and

ED visits [1]. These side effects of chemotherapy are potentially avoidable because there is a substantial institutional and geographic variation in hospital admissions and ED visits among chemotherapy patients [3] [6] [7]. A study found that among 154 gastrointestinal chemotherapy patients, 19 percent of hospitalizations were classified as potentially avoidable [2]. This shows that there is an opportunity to reduce hospitalizations among those that receive outpatient chemotherapy.

Citations

1. Aprile, G., F.E. Pisa, A. Follador, L. Foltran, F. De Pauli, M. Mazzer, S. Lutrino, C.S. Sacco, M. Mansutti, and G. Fasola. "Unplanned Presentations of Cancer Outpatients: A Retrospective Cohort Study." Supportive Care in Cancer, vol. 21, no. 2, 2013, pp. 397–404.

2. Brooks, G. A., T. A. Abrams, J. A. Meyerhardt, P. C. Enzinger, K. Sommer, C. K. Dalby, H. Uno, J. O. Jacobson, C. S. Fuchs, and D. Schrag. "Identification of Potentially Avoidable Hospitalizations in Patients with GI Cancer." Journal of Clinical Oncology, vol. 32, no. 6, 2014, pp. 496-503.

3. Brooks, G. A., L. Li, D. B. Sharma, J. C. Weeks, M. J. Hassett, K. R. Yabroff, and D. Schrag. "Regional Variation in Spending and Survival for Older Adults with Advanced Cancer." Journal of the National Cancer Institute, vol. 105, no. 9, 2013, pp. 634-642.

4. Kolodziej, M., J.R. Hoverman, J.S. Garey, J. Espirito, S. Sheth, A. Ginsburg, M.A. Neubauer, D. Patt, B. Brooks, C. White, M. Sitarik, R. Anderson, and R. Beveridge. "Benchmarks for Value in Cancer Care: An Analysis of a Large Commercial Population." Journal of Oncology Practice, vol. 7, 2011, pp. 301–306.

5. McKenzie, H., L. Hayes, K. White, K. Cox, J. Fethney, M. Boughton, and J. Dunn. "Chemotherapy Outpatients' Unplanned Presentations to Hospital: A Retrospective Study." Supportive Care in Cancer, vol. 19, no. 7, 2011, pp. 963–969.

6. Morden, N. E., C. H. Chang, J. O. Jacobson, E. M. Berke, J. P. Bynum, K. M. Murray, and D. C. Goodman. "End-of-Life Care for Medicare Beneficiaries with Cancer is Highly Intensive overall and Varies Widely." Health Affairs (Project Hope), vol. 31, no. 4, 2012, pp. 786-796.

7. Wennberg, J. E., E. S. Fisher, T. A. Stukel, J. S. Skinner, S. M. Sharp, and K. K. Bronner. "Use of Hospitals, Physician Visits, and Hospice Care during Last Six Months of Life among Cohorts Loyal to Highly Respected Hospitals in the United States." BMJ (Clinical Research Ed.), vol. 328, no. 7440, 2004, pp. 607.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

We examined associations between outcomes and SDS factors using both bivariate and multivariate analyses. On the patient-level, our analysis shows that "low SDS" patients (as characterized by three individual indicators: Medicaid dual-eligibility, race as black, and AHRQ SES Composite Index) receiving hospital-based outpatient chemotherapy are more likely to have an inpatient admission and emergency department (ED) visit within 30 days than "non-low SDS" patients.

- Dual eligible patients are more likely to have an inpatient admission or ED visit than non-dual eligible patients (13.7 percent of dual eligible vs 9.7 percent of non-dual eligible for inpatient admission, and 6.2 percent of dual eligible vs 3.8 percent of non-dual eligible for ED visits);
- Black patients are more likely to have an inpatient admission or ED visit than non-black patients (12.9 percent of black patients vs 10.0 percent of non-black for inpatient admission, and 5.5 percent of black patients vs 4.0 percent of non-black for ED visits); and
- Lower AHRQ SES Composite Index patients are more likely to have an inpatient admission or ED visit than higher SES Composite Index patients (11.5 percent of patients with low AHRQ SES Composite Index vs 9.4 percent of patients with high AHRQ SES Composite Index for inpatient admission, and 4.8 percent of patients with low AHRQ SES Composite Index vs 3.6 percent of patients with high AHRQ SES Composite Index for ED visits).

When evaluated on the hospital-level, these disparities are no longer significant. At the hospital-level, no between-hospital effects were observed for hospital case-mix by Medicaid dual-eligibility, race, or the AHRQ SES Composite Index. Specifically, there was no clear relationship between the median risk-standardized rates and hospitals' case mix by these three SDS factors. In addition, the distributions of risk-standardized rates overlapped significantly across hospitals grouping by these three SDS factors, suggesting that hospitals caring for a greater percentage of low SDS patients have similar rates of inpatient admission and ED visits within 30 days of hospital-based outpatient chemotherapy. See the NQF Testing Attachment, Section 2b4.4b and in the separate appendix titled

"ChemoMeasure_NQF Appendix_SDS" for more information on analysis and results.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Not applicable

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, A leading cause of morbidity/mortality, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
This measure is intended to assess the care provided to cancer patients and inform quality improvement efforts to reduce potentially preventable admissions and ED visits and, ultimately, health care spending. This measure addresses the National Quality Strategy priorities of promoting effective communication and coordination of care and promoting the most effective prevention and treatment practices for the leading causes of mortality. Poor performance on the measure reflects high resource use and significant consequences for patients/society due to poor quality; admissions and ED visits are costly to payers and reduce quality of life for patients.

Each year, more than a million people in the United States are diagnosed with cancer [3] with medical expenditures exceeding $100 billion [8]. According to age adjusted 2008-2012 Surveillance, Epidemiology, and End Results program data, there were 454.8 per 100,000 people diagnosed with cancer and 171.2 per 100,000 cancer-related deaths in the United States per year [3]. Approximately 22 percent of cancer patients receive chemotherapy at some point [4], with Medicare payments for cancer treatment totaling $34.4 billion in 2011 or almost 10 percent of Medicare fee-for-service (FFS) dollars [9]. Furthermore, medical expenditures for cancer are estimated to increase 27 percent from 2010 to 2020 [5]. Reducing the cost of cancer care is a high priority of the United States health care system because of the increasing annual direct costs [8].

In addressing the high cost of cancer care, it is important to focus on reducing hospitalizations because it is the single largest component of spending for cancer care [10]. In particular, unscheduled ED and hospital admissions are significant sources of utilization and cost among cancer patients [4]. On average, cancer patients receiving chemotherapy have one hospital admission and two ED visits per patient per year; approximately 40 percent of those admissions and 50 percent of ED visits are related to complications of chemotherapy [4]. Admissions and ED visits related to chemotherapy treatment pose a heavy financial burden. Using commercial claims data, Fitch and Pyenson reported that the national average cost of a chemotherapy-related admission was $22,000, and the average cost of a chemotherapy-related ED visit was $800 [1]. This measure focuses on cancer patients receiving chemotherapy treatment. Although not all cancer patients require maintenance or curative chemotherapy, focusing on this population enhances the measure in two-ways: (1) it provides an attribution methodology holding the hospital administering chemotherapy treatment responsible for care management, and (2) it assesses care at a critical time in a patient's management, when adverse events are common but can be prevented, thus having a significant impact on patient quality of life and patient outcomes.

Poor management of chemotherapy-related symptoms will lead to increased healthcare costs, morbidity, and mortality. For instance, chemotherapy patients with neutropenia are more prone to bacterial infections, a major cause of morbidity [2]. Additionally, cancer patients receiving outpatient chemotherapy who subsequently developed febrile neutropenia had higher rates of healthcare utilization than non-febrile neutropenia patients and therefore incurred higher costs [6]. Similarly, if chemotherapy-related pneumonia is not prevented or treated in a timely manner, patients can develop respiratory failure, which is costly to treat [7]. To address these issues, guidelines from various professional societies recommend evidence-based interventions to prevent and treat common side effects and complications of chemotherapy (see question 1a.2.1. in Evidence Form). Appropriate care in the outpatient setting should reduce chemotherapy-related symptoms and curb potentially avoidable hospital admissions and ED visits.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Fitch, K., and B. Pyenson. "Cancer Patients Receiving Chemotherapy: Opportunities for Better Management." Available at http://publications.milliman.com/research/health-rr/pdfs/cancer-patients-receiving-chemotherapy.pdf. Accessed on September 24, 2012.

2. Gafter-Gvili, A., A. Fraser, M. Paul, L. Vidal, T. A. Lawrie, M. D. van de Wetering, L. C. Kremer, and L. Leibovici. "Antibiotic Prophylaxis for Bacterial Infections in Afebrile Neutropenic Patients Following Chemotherapy." The Cochrane Database of Systematic Reviews, vol. 1, 2012, pp. CD004386.

3. Howlader, N., A.M. Noone, M. Krapcho, J. Garshell, D. Miller, S.F. Altekruse, C.L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D.R. Lewis, H.S. Chen, E.J. Feuer, K.A. Cronin (eds). "SEER Cancer Statistics Review, 1975-2012." National Cancer Institute. Available at http://seer.cancer.gov/csr/1975_2012/.

4. Klodziej, M., J.R. Hoverman, J.S. Garey, J. Espirito, S. Sheth, A. Ginsburg, M.A. Neubauer, D. Patt, B. Brooks, C. White, M. Sitarik, R. Anderson, and R. Beveridgel. "Benchmarks for Value in Cancer Care: An Analysis of a Large Commercial Population." Journal of Oncology Practice, vol. 7, 2011, pp. 301–306.

5. Mariotto, A. B., K. R. Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown. "Projections of the Cost of Cancer Care in the United States: 2010-2020." Journal of the National Cancer Institute, vol. 103, no. 2, 2011, pp. 117-128.

6. Michels, S.L., R.L. Barron, M.W. Reynolds, K. Smoyer Tomic, J. Yu, and G.H. Lyman. "Costs Associated with Febrile Neutropenia in the US." Pharmacoeconomics, vol. 30, no. 9, 2012, pp. 809–823.

7. Park, S.Y., S.Y. Lim, S.W. Um, W.J. Koh, M.P. Chung, H. Kim, O.J. Kwon, H.K. Park, S.J. Kim, Y.H. Im, M.J. Ahn, and G.Y. Suh. "Outcome and Predictors of Mortality in Patients Requiring Invasive Mechanical Ventilation Due To Acute Respiratory Failure While Undergoing Ambulatory Chemotherapy for Solid Cancers." Supportive Care in Cancer, vol. 21, no. 6, 2013, pp. 1647–1653.

8. Smith, T. J. and B. E. Hillner. "Bending the Cost Curve in Cancer Care." The New England Journal of Medicine, vol. 364, no. 21, 2011, pp. 2060-2065.

9. Sockdale, H., K. Guillory. "Lifeline: Why Cancer Patients Depend on Medicare for Critical Coverage." Available at http://www.acscan.org/content/wp-content/uploads/2013/06/2013-Medicare-Chartbook-Online-Version.pdf.

10. Yabroff, K. R., E. B. Lamont, A. Mariotto, J. L. Warren, M. Topor, A. Meekins, and M. L. Brown. "Cost of Care for Elderly Cancer Patients in the United States." Journal of the National Cancer Institute, vol. 100, no. 9, 2008, pp. 630-641.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable

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 subcriteria 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):*

Cancer

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Care Coordination, Safety : Complications, Safety : Medication Safety

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.)*
The following webpage contains a copy of the full technical report for this measure: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.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:** ChemoMeasure_NQF_Attachment_Data_Dictionary.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not applicable

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
This measure involves calculating two mutually exclusive outcomes: one or more inpatient admissions or one or more ED visits for any of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of chemotherapy treatment among cancer patients receiving treatment in a hospital outpatient setting. These 10 conditions are potentially preventable through appropriately managed outpatient care. The qualifying diagnosis on the admission or ED visit claim must be (1) the principal diagnosis or (2) a secondary diagnosis accompanied by a principal diagnosis of cancer.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator (outcome) time frame: 30-day period (inclusive) following the date of each chemotherapy treatment in an outpatient setting at the reporting hospital.

Denominator (cohort) time frame: Any chemotherapy treatment performed in a hospital outpatient setting during the performance period (e.g., 1 year).

Risk adjustment look-back period: 1 year prior to date of first chemotherapy treatment during the performance period.

S.6. **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, 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.
Outcome Definition:
This measure has two reported outcomes. The outcomes for this measure are one or more inpatient admissions or one or more ED visits for one of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of receiving hospital outpatient chemotherapy treatment for cancer. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes that identify these diagnoses are in the attached Data Dictionary on sheets "S.6 Numerator-Anemia," "S.6 Numerator-Dehydration," "S.6 Numerator-Diarrhea," "S.6 Numerator-Emesis," "S.6 Numerator-Fever," "S.6 Numerator-Nausea," "S.6 Numerator-Neutropenia," "S.6 Numerator-Pain," "S.6 Numerator-Pneumonia," and "S.6 Numerator-Sepsis." The ICD-9-CM codes were used during development and testing; the Data Dictionary also includes the mapping from these ICD-9-CM codes to ICD-10-CM codes.

Identification of Outcomes:
Outcomes are identified using Medicare Part A Inpatient and Part B Outpatient hospital claims. The qualifying diagnosis on the

admission or ED visit claim must be listed as (1) the principal diagnosis or (2) a secondary diagnosis accompanied by a principal diagnosis of cancer. These ten conditions are considered potentially preventable through appropriately managed outpatient care. Outcomes are identified separately for the inpatient and ED categories. A patient can only qualify for an outcome once. Patients who experience both an inpatient admission and an ED visit during the performance period are counted towards the inpatient admission outcome. Among those with no qualifying inpatient admissions, qualifying ED visits will be counted. As a result, the rates can be viewed as additive to provide a comprehensive performance estimate of quality of care following hospital-based outpatient chemotherapy treatment. The rates are calculated separately because the severity and cost of an inpatient admission is different from that of an ED visit, but both adverse events are important signals of quality and represent patient-important outcomes of care.

Outcome attribution:
The measure attributes the outcome to the hospital where the patient received chemotherapy treatment during the 30 days before the outcome. If a patient received outpatient chemotherapy treatment from more than one hospital in the 30 days before the outcome, the measure will attribute the outcome to all the hospitals that provided treatment in those 30 days. For example, if a patient received an outpatient chemotherapy treatment at Hospital A on January 1, a second treatment at Hospital B on January 10, and then experienced a qualifying admission on January 15, the measure would count this outcome for both Hospital A and Hospital B because both hospitals provided outpatient chemotherapy treatment to the patient within the 30-day window. However, if a patient received an outpatient chemotherapy treatment from Hospital A on January 1, and a second treatment from Hospital B on March 1, and then experienced a qualifying outcome on March 3, the measure would attribute this outcome only to Hospital B. Note that in the testing of this measure, using Medicare Fee-For-Service (FFS) claims data from July 1, 2012, to June 30, 2013, only 5 percent of patients in the cohort received outpatient chemotherapy treatment from more than one facility during the year.

Outcome Time Frame:
The measure limits the outcome time frame to the 30 days following the date of each chemotherapy treatment (including the day of treatment) in an outpatient setting for four reasons. First, existing literature suggests the vast majority of adverse events occur within 30 days after treatment [1, 2, 3], indicating that a 30-day period is a reasonable timeframe to observe the side effects of treatment. Second, we observed in our own data that the highest rates of hospital visits occur within 30 days after chemotherapy treatment. Third, restricting the time period ensures that patients' experiences are attributed to the hospitals that provided their recent treatment while accounting for variations in duration between outpatient treatments. Fourth, relating the time frame to a specific chemotherapy administration supports the idea that the admission stems from the management of side effects of treatment and ongoing care, rather than progression of the disease or other unrelated events.

Citations:
1. Aprile, G., F.E. Pisa, A. Follador, L. Foltran, F. De Pauli, M. Mazzer, S. Lutrino, C.S. Sacco, M. Mansutti, and G. Fasola. "Unplanned Presentations of Cancer Outpatients: A Retrospective Cohort Study." Supportive Care in Cancer, vol. 21, no. 2, 2013, pp. 397–404.
2. Foltran, L., G. Aprile, F.E. Pisa, P. Ermacora, N. Pella, E. Iaiza, E. Poletto, S.E. Lutrino, M. Mazzer, M. Giovannoni, G.G. Cardellino, F. Puglisi, and G. Fasola. "Risk of Unplanned Visits for Colorectal Cancer Outpatients Receiving Chemotherapy: A Case-Crossover Study." Supportive Care in Cancer, vol. 22, no. 9, 2014, pp. 2527–2533.
3. McKenzie, H., L. Hayes, K. White, K. Cox, J. Fethney, M. Boughton, and J. Dunn. "Chemotherapy Outpatients' Unplanned Presentations to Hospital: A Retrospective Study." Supportive Care in Cancer, vol. 19, no. 7, 2011, pp. 963–969.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
The measure cohort includes Medicare FFS patients aged 18 years and older as of the start of the performance period with a diagnosis of any cancer who received at least one hospital outpatient chemotherapy treatment at the reporting hospital during the performance period.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Target Population:
The target population is patients aged 18 years and older who are enrolled in Medicare FFS with a diagnosis of cancer (except leukemia; see Denominator Exclusion, Section S.10 for exclusion details) during the performance period and at least one chemotherapy treatment performed in a hospital outpatient department (OPD). The ICD-9 codes that identify cancer diagnoses are in the attached Data Dictionary, sheets "S.9 Denominator-Cancer." The ICD-9-CM codes were used during development and testing;

the Data Dictionary also includes the mapping from these ICD-9-CM codes to ICD-10-CM codes.

Target Age Group:
This measure includes all adult Medicare FFS patients because all adult patients with a treatment plan allowing for chemotherapy treatment in a hospital outpatient setting should have their care properly managed to reduce the need for acute care for the specific conditions on which this measure focuses. Additionally, by including all adult patients, rather than restricting to those 65 years of age or older, the measure assesses a broader population and more comprehensively evaluates the quality of care provided by the hospital OPD.

Because persons under 65 enrolled in Medicare and with a cancer diagnosis are likely to differ substantially from the older population with cancer due to either their co-existing medical conditions and/or the cancer stage, we explored the appropriateness of including Medicare patients aged 18 to 64 years in the cohort by (1) reviewing patient characteristics separately for these two subsets, (2) reviewing the observed performance rates for the two subsets, and (3) fitting the risk adjustment model separately for these subsets. We found that patients aged 18 to 64 years represent 13 percent of the final measure cohort, and although the younger population has higher observed outcome rates, the inpatient admission and ED visit risk-adjustment models behave similarly on both subsets of patients (patients 18-64 and patients 65+). The risk-adjustment models fit both subsets of patients similarly because the models capture and adjust for key differences, such as age, cancer type, and comorbidities, which are likely to vary between the two groups (see results in the attached Data Dictionary, sheet “S.15 Risk Model Specs” for risk factors). Based on these findings, we determined there was not a strong statistical or clinical reason to exclude the younger patients. We therefore include all adult patients 18 years and older in the measure cohort.

Focus on Chemotherapy Provided in Outpatient Setting:
This measure focuses on the management of symptoms for patients receiving care in the hospital outpatient setting and is not intended to be a comprehensive assessment of the level of symptom management for all cancer patients treated at the hospital (inpatient and outpatient). Rather, this measure assesses an aspect of care with documented unmet patient needs resulting in reduction of patient´s quality of life and increase in healthcare utilization and costs. Several studies illustrate a gap in care for outpatients as they are "invisible" from the system when they return home following treatment [1, 3, 4]. In addition, this measure focuses on treatments in the hospital outpatient setting rather than in the inpatient setting because of the increase in hospital-based chemotherapy, which presents an opportunity to coordinate care. Among Medicare patients who are receiving chemotherapy, from 2008 to 2012 the proportion of those patients receiving chemotherapy in a hospital-based outpatient setting (as opposed to a physician office) increased from 18 to 29 percent, and this trend is likely to continue [5]. By focusing the measure on this population, we think the performance rate provides meaningful and actionable feedback to hospitals.

Identifying Chemotherapy Patients in the Hospital Outpatient Setting:
During development we considered the most appropriate target population—all chemotherapy patients or limit to only patients on palliative treatment regimens, where keeping patients out of the hospital is a desirable outcome and focus of care improvement. We worked with a range of stakeholders, including oncologists and cancer center and hospital representatives, throughout the measure development process to reach consensus on the measure intent and specifications. Through these efforts and review of published literature, we have determined that all patients receiving outpatient chemotherapy, regardless of the reason for chemotherapy (palliative vs curative) may experience a gap in care that leads to acute, potentially preventable hospitalizations, and that improving patients´ quality of life by keeping patients out of the hospital is a main goal of cancer care. As a result, this measure currently includes all patients receiving chemotherapy in the hospital OPD and focuses on preventable reasons for admission. Regardless of the reason for chemotherapy, providers should assess patient risks and take preventative action where possible; communication lines should be open so the patient clearly understands expectations and how to handle. Additionally, the reason for treatment cannot be determined from claims data.

The measure identifies chemotherapy treatment using ICD-9-CM procedure and encounter codes and Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS) procedure and medication procedure codes. The ICD-9-CM, CPT, and HCPCS codes that identify chemotherapy treatment are in the attached Data Dictionary, sheets “S.9 Denominator-Chemo Procedure,” “S.9 Denominator-Chemo Encounter,” and “S.9 Denominator-Chemo Medicine.” The measure excludes procedure codes for oral chemotherapy because it is challenging to identify oral chemotherapy without using pharmacy claims data and, according to our TEP, most oral chemotherapies have fewer adverse reactions that result in admissions.

We have developed a ‘coding crosswalk’ between ICD-9-CM codes and ICD-10-CM codes. For detailed information on the cohort definition including the ICD-9-CM, ICD-10-CM, CPT, and HCPCS codes that identify chemotherapy treatment, see the Data Dictionary appendix.

Inclusion of Chemotherapy Treatments Affected by the Medicare 3-Day Payment Window Policy:
The measure depends on identifying chemotherapy treatments performed in hospital OPDs. The Medicare 3-day payment window affects our ability to identify some outpatient chemotherapy treatments performed in hospital OPDs that resulted in an admission. The policy states that outpatient services (including some non-diagnostic services such as chemotherapy) provided by a hospital or any Part B entity wholly owned or wholly operated by a hospital (such as a hospital OPD) in the three calendar days preceding the date of a beneficiary's inpatient admission are deemed to be related to the admission [2]. For outpatient chemotherapy treatments subject to the 3-day payment policy, the outpatient chemotherapy service should be bundled and billed with the inpatient claim.

To ensure the inclusion of all hospital OPD chemotherapies, the measure first identifies all chemotherapy treatments during the performance period within the hospital outpatient claims file and then supplements this cohort by identifying chemotherapy treatments included on inpatient claims with a date of service prior to or equal to the date of admission on the claim. The measure includes outpatient-based chemotherapy procedures on inpatient claims with the same date of service as the admission date because, clinically, patients would receive an outpatient chemotherapy treatment and then have a qualifying inpatient admission. That is, we do not expect cancer patients with a qualifying admission for the 10 potentially preventable conditions to receive chemotherapy on that same day, as generally they would not receive chemotherapy if they required acute care for these diagnoses. Moreover, the expectation is that chemotherapy administration and the surrounding care is what accounted for the qualifying diagnosis that was the principal reason for the admission or ED visit. We will continue to assess this approach to identifying chemotherapy treatments subject to CMS 3-day payment window billing during annual measure maintenance and prior to implementation.

Citations:
1. Aprile, G., F.E. Pisa, A. Follador, L. Foltran, F. De Pauli, M. Mazzer, S. Lutrino, C.S. Sacco, M. Mansutti, and G. Fasola. "Unplanned Presentations of Cancer Outpatients: A Retrospective Cohort Study." Supportive Care in Cancer, vol. 21, no. 2, 2013, pp. 397–404.

2. Centers for Medicare & Medicaid Services (CMS). Three Day Payment Window. 2013; http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Three_Day_Payment_Window.html

3. Foltran, L., G. Aprile, F.E. Pisa, P. Ermacora, N. Pella, E. Iaiza, E. Poletto, S.E. Lutrino, M. Mazzer, M. Giovannoni, G.G. Cardellino, F. Puglisi, and G. Fasola. "Risk of Unplanned Visits for Colorectal Cancer Outpatients Receiving Chemotherapy: A Case-Crossover Study." Supportive Care in Cancer, vol. 22, no. 9, 2014, pp. 2527–2533.

4. McKenzie, H., L. Hayes, K. White, K. Cox, J. Fethney, M. Boughton, and J. Dunn. "Chemotherapy Outpatients' Unplanned Presentations to Hospital: A Retrospective Study." Supportive Care in Cancer, vol. 19, no. 7, 2011, pp. 963–969.

5. Vandervelde, Aaron, Henry Miller, and JoAnna Younts. "Impact on Medicare Payments of Shift in Site of Care for Chemotherapy Administration." Washington, DC: Berkeley Research Group, June 2014. Available at http://www.communityoncology.org/UserFiles/BRG_340B_SiteofCare_ReportF_6-9-14.pdf. Accessed September 16, 2015.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
We established the following exclusion criteria after reviewing the literature, examining existing measures, reviewing feedback from a public comment period, and discussing alternatives with the Cancer Working Group and TEP members (see Section Ad.1. for description of group and membership). The goal was to be as inclusive as possible; we excluded only those patient groups for which hospital visits were not typically a quality signal or for which risk adjustment would not be adequate. The exclusions, based on clinical rationales, prevent unfair distortion of performance results.

1) Patients with a diagnosis of leukemia at any time during the performance period.

Rationale: Patients with leukemia are excluded due to the high toxicity of treatment and recurrence of disease so that admissions do not reflect poorly managed outpatient care for this population. Patients with leukemia have an expected admission rate due to relapse, so including leukemia patients in the cohort could be conceptualized as a planned admission, which does not align with the intent of the measure.

2) Patients who were not enrolled in Medicare FFS Parts A and B in the year prior to the first outpatient chemotherapy treatment during the performance period.

Rationale: We exclude these patients to ensure complete patient diagnosis data for the risk-adjustment model, which uses the year prior to the first chemotherapy treatment during the period to identify comorbidities.

3) Patients who do not have at least one outpatient chemotherapy treatment followed by continuous enrollment in Medicare FFS Parts A and B in the 30 days after the procedure.

Rationale: We exclude these patients to ensure full data availability for outcome assessment.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)*
1) Patients with a diagnosis of leukemia at any time during the performance period.

Details: The ICD-9-CM codes that define leukemia are in the attached Data Dictionary, sheet "S.11 Denom Exclusion-Leukemia." We check hospital inpatient, hospital outpatient, and Carrier (Part B) claims for a diagnosis of leukemia. The ICD-9-CM codes were used during development and testing; the Data Dictionary also includes the mapping from these ICD-9-CM codes to ICD-10-CM codes.

2) Patients who were not enrolled in Medicare FFS Parts A and B in the year prior to the first outpatient chemotherapy treatment during the performance period.

Details: Lack of continuous enrollment in Medicare FFS for the 12 months prior to the first procedure during the performance period is determined by patient enrollment status in FFS Parts A and B using the Medicare enrollment files. The enrollment indicators must be appropriately marked for all 12 months which fall within 1 year prior to the procedure date.

3) Patients who do not have at least one outpatient chemotherapy treatment followed by continuous enrollment in Medicare FFS Parts A and B in the 30 days after the procedure.

Details: Lack of continuous enrollment in Medicare FFS for 1 month after the procedure is determined by patient enrollment status in FFS Parts A and B using the Medicare enrollment files. The enrollment indicators must be appropriately marked for the month(s) which falls within 30 days of the procedure date.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not applicable. This measure is not stratified.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Our approach to risk adjustment is tailored to, and appropriate for, a publicly reported outcome measure as articulated in published scientific guidelines [1,2].

Since the measure has two mutually exclusive outcomes—qualifying inpatient admissions and qualifying ED visits—we developed two risk-adjustment models, one for each dependent variable (inpatient admissions and ED visits). We use a two-level hierarchical logistic regression model to estimate risk-standardized outcome rates. This approach accounts for differences in patient mix, the clustering of patients within hospitals, and variation in sample size.

The measure adjusts for variables that are clinically relevant and associated with the outcome. It seeks to adjust for differences in patient demographics, clinical comorbidities, and treatment exposure (that is, the number of chemotherapy treatments undergone by the patient in the hospital outpatient setting during the performance period), which vary across patient populations and influence the outcome but do not relate to quality. Specifically, the risk-standardization model for inpatient admissions has 20 patient-level variables (age, sex, exposure, nine comorbidity variables, and eight cancer categories). The risk-standardization model for ED visits has 15 patient-level variables (age, sex, exposure, six comorbidity variables, and six cancer categories). We define the exposure variable as the count of hospital outpatient chemotherapy treatments the patient received during the performance period; it is important to adjust for exposure because the more treatments a patient receives, the higher the chance that one of

them will be followed by a qualifying outcome.

We define comorbidity variables using condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9 diagnosis codes. A map showing the assignment of ICD-9-CM codes to CCs can be found in the attached Data Dictionary, sheet "S.14 ICD-9 to CC mapping." We worked with a subset of our TEP to narrow the CCs to those most appropriate for this measure. In reviewing the CCs to identify those conditions appropriate for inclusion in our model, we considered the number of patients potentially affected, whether the condition affects admission for one of the ten outcome qualifying diagnoses, and whether inclusion of the condition in the model would incentivize appropriate treatment.

The end result was 9 bundled CCs for potential inclusion in the models: (1) diabetes, (2) metabolic disorders, (3) gastrointestinal (GI) disorders, (4) psychiatric disorders, (5) neurological conditions, (6) cardiovascular disease, (7) respiratory disorders, (8) renal disease, and (9) other injuries. We define the cancer type in nine categories developed based on clinical similarities and distribution of patients. The nine categories for potential inclusion in the model included: (1) breast cancer, (2) digestive cancer, (3) genitourinary cancer, (4) respiratory cancer, (5) lymphoma, (6) prostate cancer, (7) secondary cancer of the lymph nodes, (8) secondary cancer of solid tumors, and (9) other cancers. The Condition Categories (CCs) that define each of these comorbidities and the ICD-9-CM codes that define the cancer categories are included in the Data Dictionary, on the following sheets "S.15 Risk Model Specs."

Inpatient Admission Model Variables
The patient-level risk-adjustment variables are:
1. Age (continuous)
2. Sex (male)
3. Exposure
4. Respiratory Disorder (CC 107-110)
5. Renal Disease (CC 128-131)
6. Diabetes (CC 15-20)
7. Other Injuries (CC 162)
8. Metabolic Disorder (CC 21-24)
9. Gastrointestinal Disorder (CC 25-36)
10. Psychiatric Disorder (CC 48-66)
11. Neurological Conditions (CC 67-76)
12. Cardiovascular Disease (CC 77-106)
13. Breast Cancer
14. Digestive Cancer
15. Respiratory Cancer
16. Lymphoma
17. Prostate Cancer
18. Secondary Cancer of Lymph Nodes
19. Secondary Cancer of Solid Tumors
20. Other Cancer

ED Visits Model Variables
The patient-level risk-adjustment variables are:
1. Age (continuous)
2. Sex (male)
3. Exposure
4. Respiratory Disorder (CC 107-110)
5. Other Injuries (CC 162)
6. Gastrointestinal Disorder (CC 25-36)
7. Psychiatric Disorder (CC 48-66)
8. Neurological Conditions (CC 67-76)
9. Cardiovascular Disease (CC 77-106)
10. Breast Cancer
11. Digestive Cancer
12. Respiratory Cancer
13. Secondary Cancer of Lymph Nodes
14. Secondary Cancer of Solid Tumors
15. Other Cancer

Citations

1. Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. 2006; 113 (3): 456-462.

2. Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci. 2007; 22 (2): 206-226.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
The measure estimates hospital-specific risk-adjusted rates of potentially avoidable inpatient admissions or ED visits for cancer patients aged 18 years or older receiving chemotherapy treatment in a hospital OPD using a hierarchical logistic regression model. The cohort includes Medicare FFS patients aged 18 years or older at the start of the performance period with a diagnosis of cancer (other than leukemia) during the performance period who: had at least one hospital outpatient chemotherapy treatment during the performance period; were enrolled in Part A and Part B Medicare for the 12 months prior to the first chemotherapy treatment during the performance period; and were enrolled in Part A and B for the 30 days following at least one outpatient chemotherapy treatment. A single patient may be attributed to more than one hospital if the patient received chemotherapy treatment in a hospital OPD from more than one hospital during the performance period.

For each patient in the cohort, two outcomes are assessed. The first outcome is defined as any inpatient admissions within 30 days of any chemotherapy treatment in a hospital OPD during the performance period with either (a) a principal diagnosis of anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis or (b) a principal diagnosis of cancer and one of those ten diagnoses listed as secondary on the same claim. These ten conditions are seen as potentially preventable through appropriately managed outpatient care. The second outcome is defined as ED visits over that same time period with the same qualifying diagnoses. The second outcome is assessed only for patients who do not qualify for the first outcome. In addition, a patient can only qualify for an outcome once. As a result, the rates can be viewed as additive to get a full picture of patients in the cohort that had at least one potentially preventable outcome. The rates are calculated separately because severity and cost of inpatient admission is different from an ED visit, but both are adverse events and important pieces of information for quality improvement efforts.

These rates are risk-adjusted using hierarchical regression models; separate models are utilized for each outcome. The measure calculates the hospital-specific risk-adjusted rate as the ratio of a hospital's "predicted" number of outcomes to "expected" number of outcomes multiplied by the national observed outcome rate. It estimates the expected number of outcomes for each hospital using the hospital's patient mix and the average hospital-specific intercept (that is, the average intercept among all hospitals in the sample). The measure estimates the predicted number of outcomes for each hospital using the same patient mix, but an estimated hospital-specific intercept. Operationally, the measure obtains the expected number of outcomes for each hospital by summing the expected probabilities of outcomes for all patients treated at the hospital. It calculates the expected probability of outcomes for each patient via the hierarchical model, which applies the estimated regression coefficients to the observed patient characteristics

and adds the average of the hospital-specific intercept. It calculates the predicted number of outcomes for each hospital by summing the predicted probabilities for all patients in the hospital. The measure calculates the predicted probability for each patient through the hierarchical model, which applies the estimated regression coefficients to the observed patient characteristics and adds the hospital-specific intercept.

If a hospital's ratio of predicted to expected outcomes is less than 1, it indicates that the hospital is performing better than expected given its case mix. If a hospital's ratio of predicted to expected outcomes is greater than 1, it indicates that the hospital is performing worse than expected given its case mix. For ease of interpretation, we transform this ratio to a rate by multiplying by the national observed rate for that outcome. If the "predicted" number of outcomes is higher (or lower) than the "expected" number of outcomes for a given hospital, the risk-adjusted rate will be higher (or lower) than the national observed admission rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample or survey.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This measure is not based on a sample or survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not applicable

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Medicare administrative claims and enrollment data.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **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

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
ChemoMeasure_NQF_testing_attachment.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Admissions and Emergency Department (ED) Visits for Patients Receiving Outpatient Chemotherapy

Date of Submission: 3/11/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decision making) 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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.
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.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*).

To develop and test the measure, the team used 2012-2013 Medicare 100% Fee-For-Service (FFS) data.

We used the following datasets to define the cohort and collect data for risk adjustment:

- Medicare hospital outpatient Standard Analytic File (SAF): To capture chemotherapy treatment administered in a hospital outpatient department and to identify cancer diagnoses and comorbidities
- Medicare hospital inpatient SAF: To capture chemotherapy treatment administered in a hospital outpatient department that may be bundled on an inpatient claim due to the CMS 3-day payment window policy and to identify cancer diagnoses and comorbidities
- Carrier (Part B Physician) claims SAF: To identify cancer diagnoses and comorbidities
- Medicare Enrollment Database and denominator files: To determine enrollment and demographic information

We used the following datasets to define the outcome:

- Medicare hospital outpatient SAF: To identify qualifying emergency department visits
- Medicare hospital inpatient SAF: To identify qualifying hospital admissions

In addition to the patient-level claims, we linked the patient-level claims data with a 5-digit zip-code level dataset for SDS assessment (SDS assessment is described in Sections 1.8, 2b4.3, and 24b.4b of this form and in the separate appendix titled "ChemoMeasure_NQF Appendix_SDS").

Similar to the approach used for other CMS outcome measures (for example, CMS 30-day AMI readmission measure (NQF measure #0505) developed by CMS available at: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80367), we linked the claims data with data from the American Community Survey (ACS) by patient 5-digit zip code. We used the Agency for Healthcare Research and Quality (AHRQ)-validated composite index of SES which has been used and tested among Medicare beneficiaries to create a "neighborhood-level" SES composite index. See Section 1.8 of this form for more specifics on the variables used.

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

We used data from 07/01/2012 through 06/30/2013.

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ 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*)

For this measure, hospital outpatient departments are the measured entities. The number of measured entities (hospital outpatient departments) varies by testing type; see Section 1.7 for details.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)?
(*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
The number of patients varies by testing type; see Section 1.7 for details.

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.

We used a testing cohort created from Medicare data described in Section 1.2 to develop and test aspects of this measure. We applied the measure specifications to each dataset to develop the testing cohort (see the Measure Submission Form, Sections S.4 to S.11, for full inclusion and exclusion criteria). The information below represents the final cohort, after measure exclusions are applied.

We used three datasets: (1) July 2012 – June 2013 Full Sample, (2) Development Split Sample, derived from the Full Sample, and (3) Validation Split Sample, derived from the Full Sample. The datasets, dates, number of measured entities, number of patients, and demographic profile for the patients used in each type of testing are as follows:

<u>(1) 2012-2013 Full Sample</u>

- **Dates:** July 1, 2012 - June 30, 2013
- **Number of hospitals:** 3,765
- **Number of patients and patient characteristics:**
    - 240,446 unique patients qualified in the cohort after exclusions (252,408 total patient-provider combinations, as 5% of the patients have an outpatient chemotherapy treatment at more than one hospital)
        - Age: Average age of 72.2 years
        - Sex: 50.2% male
        - Exposure: On average, patients received 5.4 chemotherapy treatments during the performance period. (median = 3, 25th percentile = 2, 75th percentile = 7)
        - Cancer type: The top three primary cancer types were Digestive Cancer (24.2%), Respiratory Cancer (21.8%), and Genitourinary Cancer (19.8%). Additionally, 39.8 percent of the cohort were identified as having "Other Cancers." All cancer types are defined in the Data Dictionary on sheets, "S.15 RM-Breast Cancer" through "S.15 RM-Sec Neo of Solid Tumor."
- **Dataset used for:** defining the cohort, testing the exclusion criteria, testing measure level reliability, and analyses to address potential threats to validity (see Sections 2a2, 2b2, and 2b3)

The 2012-2013 Development and Validation Split Samples were derived by selecting two random samples without replacement from the Medicare 2012-2013 Full Sample. Each patient-provider combination had equal probability of selection into either the development or the validation sample.

<u>(2) 2012-2013 Development Split Sample</u>

- **Dates:** July 1, 2012 - June 30, 2013 (half of the 2012-13 Full Sample)
- **Number of hospitals:** 3,483
- **Number of patients and patient characteristics:**
    - 123,149 unique patients qualify for cohort after exclusions (126,204 total patient-provider combinations)
        - Age: Average age of 72.2 years
        - Sex: Patients were evenly divided with 50.1% male
        - Exposure: On average, a patient received 5.2 chemotherapy treatments during the performance period. (median = 3, 25th percentile = 1, 75th percentile = 7)

        - Cancer type: The top three primary cancer types were Digestive Cancer (24.3%), Respiratory Cancer (21.8%), and Genitourinary Cancer (19.7%). Additionally, 28.2 percent of the cohort were identified as having "Other Cancers." All cancer types are defined in the Data Dictionary on sheets, "S.15 RM-Breast Cancer" through "S.15 RM-Sec Neo of Solid Tumor."
- **Dataset used for:** testing data element reliability and testing the patient-level risk-adjustment model (see Sections 2b4.5-2b4.7)

(3) 2012-2013 Validation Split Sample

- **Dates:** July 1, 2012 - June 30, 2013 (remaining half of the 2012-13 Full Sample)
- **Number of hospitals:** 3,469
- **Number of patients and patient characteristics:**
    - Count: 123,115 unique patients qualify for cohort after exclusions (126,204 total patient-provider combinations)
        - Age: Average age of 72.1 years
        - Sex: Patients were fairly evenly divided with 50.0% male
        - Exposure: On average, a patient received 5.2 chemotherapy treatments during the performance period. (median = 3, 25th percentile = 1, 75th percentile = 7)
        - Cancer type: The top three primary cancer types were Digestive Cancer (24.1%), Respiratory Cancer (21.7%), and Genitourinary Cancer (20.0%). Additionally, 28.3 percent of the cohort were identified as having "Other Cancers." All cancer types are defined in the Data Dictionary on sheets, "S.15 RM-Breast Cancer" through "S.15 RM-Sec Neo of Solid Tumor."
- **Dataset used for:** testing data element reliability and testing the patient-level risk-adjustment model (see Sections 2b4.5-2b4.7)

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We considered SDS factors that align with the NQF guidelines for developers and can be linked to claims data. The NQF-convened Expert Panel that considered risk-adjustment for SDS recognized risk adjustment may be constrained by data limitations and data collection burden [5]. The variables that are available within, or that can be linked directly, to Medicare administrative claims data used for this measure include those listed below. A more comprehensive explanation of considerations that went into variable selection is provided in the separate appendix titled "ChemoMeasure_NQF Appendix_SDS."

Race (black, other)

Data source: Medicare enrollment database

The particular case of race as a predictor of health outcomes illuminates the complexity of the role SDS variables play in assessing hospital performance. The association between patient race and high symptom burden has been observed among cancer patients [2,4]. Those that are undertreated for their symptoms might be more likely to seek care in the ED or be admitted to the hospital compared to those with adequate symptom management. It is possible this association may also confound with other SDS factors such as SES and geographic access to care.

Medicaid dual-eligible status (Medicaid-Medicare dual, Medicare only)

Data source: Medicare enrollment database

The dual-status patient-level variable provides a reliably-obtained indication of patients with low income/assets and high health care spending. Income is an important SDS factor to consider for this measure because it may reflect access to resources, ability to purchase medications to manage symptoms, adherence to scheduled follow-up appointments for routine check-ins and timely care, availability of family support, and more.

Neighborhood SES factors as proxies for patient-level SES [1]

Data source: Enrollment database and Census data (American Community Survey)

The American Community Survey (ACS) provides a number of SDS indicators that are available at the ZIP code level and can be linked directly to Medicare claims at the 5-digit ZIP code level. We used the Agency for Healthcare Research and Quality (AHRQ)-validated composite index of SES which has been used and tested among Medicare beneficiaries [1]. This index is a composite of seven different variables found in the Census data which may capture SDS better than any single variable. The variables are: (1) median household income, (2) percentage of persons living below the federal poverty level, (3) percentage of persons who are aged >16 years and in the labor force but not employed, (4) median value of owner-occupied homes, (5) percentage of persons aged >25 years who completed at least a 12th grade education, (6) percentage of persons aged >25 years who completed at least four years of college, and (7) percentage of households that average one or more persons per room.

This neighborhood-level variable, which we use as a proxy for patient-level SDS factors, are important to consider for this measure because they may reflect patient income (discussed above), patient's health literacy level, which is associated with higher ED use [3] [6], and home environment, where outpatients handle their care out of sight of the hospital.

References:

1. Creation of New Race-Ethnicity Codes and Socioeconomic Status (SES) Indicators for Medicare Beneficiaries: Final Report. August 2012. Agency for Healthcare Research and Quality, Rockville, MD. http://archive.ahrq.gov/research/findings/final-reports/medicareindicators/index.html. Accessed August 24, 2015.

2. Fisch, M.J., J.W. Lee, M. Weiss, L.I. Wagner, V.T. Chang, D. Cella, J.B. Manola, L.M. Minasian, W. McCaskill-Stevens, T.R. Mendoza, and C.S. Cleeland. "Prospective, Observational Study of Pain and Analgesic Prescribing in Medical Oncology Outpatients with Breast, Colorectal, Lung, or Prostate Cancer." *Journal of Clinical Oncology*, vol. 30, no. 16, 2012, pp. 1980–1988.

3. Griffey, Richard T., Sarah K. Kennedy, Lucy McGownan, Melody Goodman, and Kimberly A. Kaphingst. "Is Low Health Literacy Associated with Increased Emergency Department Utilization and Recidivism?" *Academic Emergency Medicine,*vol. 21, no. 10, 2014, pp. 1109-1115.

4. Miaskowski, C., B. A. Cooper, M. Melisko, L. M. Chen, J. Mastick, C. West, S. M. Paul, L. B. Dunn, B. L. Schmidt, M. Hammer, F. Cartwright, F. Wright, D. J. Langford, K. Lee, and B. E. Aouizerat. "Disease and Treatment Characteristics do Not Predict Symptom Occurrence Profiles in Oncology Outpatients Receiving Chemotherapy." *Cancer,* vol. 120, no. 15, 2014, pp. 2371-2378.

5. National Quality Forum. "Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report." August 15, 2014. http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx

6. Schumacher, J. R., A. G. Hall, T. C. Davis, C. L. Arnold, R. D. Bennett, M. S. Wolf, and D. L. Carden. "Potentially Preventable use of Emergency Services: The Role of Low Health Literacy." *Medical Care,* vol. 51, no. 8, 2013, pp. 654-658.

_____

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*)

Data Element Reliability

In constructing the measure in Medicare FFS patients, we aim to utilize only those data elements from claims data that have both face validity and reliability. Specifically, we use fields that are consequential for payment and which are audited. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and recoup overpayment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard.

Measure Score Reliability

The reliability of a measurement refers to the degree to which repeated measurements of the same entity agree with each other. Specifically, for hospital-level performance measures, reliability characterizes to what extent repeated measurements of the same hospital generate similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we assessed measure reliability for the measure using split-half correlations with claims data for all hospitals. To calculate split-half reliability, we randomly divided the hospital-level data into two equal samples; thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. We calculated the measure performance in both samples for each hospital; to the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the Pearson correlation between the performance rate estimates: the higher the correlation, the higher the reliability of the measure. In order to produce estimates that are as stable as possible, we repeated this approach 1,000 times, a technique known as bootstrapping [1].

Because we expect hospitals with relatively few cases to have less reliable estimates, we only included scores for hospitals with at least 60 patients in the reliability calculation (i.e., with 30 patients in each of the split samples). This approach is consistent with a reporting strategy that includes smaller hospitals in the measure calculation, but does not publicly release the measure score for smaller hospitals (i.e., labels them in public reporting as having "too few cases" to support a reliable estimate). We note that the minimum sample size for public reporting is a policy choice that balances competing considerations such as the reliability of the measure score and transparency for consumers, and that the cutoff used for this analysis is one of many that might be reasonably used.

In addition, we conducted a second analysis of measure reliability using the intraclass correlation coefficient (ICC) signal-to-noise method to determine a recommended minimum case count to maintain a moderate level of reliability. The ICC

is estimated from the random effects model that produces the risk-standardized hospital visit rates, as ICC = V / (V + σ), where V is the between variance and σ is the sampling variance of the estimated provider level results. Because $\pi^2/3$ is the sampling variance of the logit distribution, the ICC of the measure, which is based on a logit model, is ICC = V / (V + $\pi^2/3$).

We used the intercept variance from the hierarchical logit models used to estimate the measure (0.0909 for inpatient admission, and 0.1108 for ED visits) as the estimate of the between variance. The ICC can be used to calculate the reliability (R) of individual hospitals using the formula: R= N/(N + (1 - ICC)/ICC).[1] The case size required for a given R is: N = R*(1 - ICC)/(ICC*(1 - R)). We looked for the N required to maintain a reliability level of 0.4 or higher.

Citations

1. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002; 21:3431-3446.

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*)

Data Element Reliability

We did not conduct statistical analyses at the data element level. We did review patient characteristics and frequency of risk variables across our cohort with the TEP and other clinicians for reasonability. Some of this information is presented in Section 1.7, and complete information is included in Tables 1 and 2 the attached Measure Technical Report.

Measure Score reliability

There were 942 hospitals with ≥60 patients in their cohorts in the full sample. This sample was randomly split 1,000 times and the Pearson correlation was calculated each time. For the inpatient admission measure, on average, the agreement between the two hospital visit rates for each hospital was 0.413 (95% confidence interval (CI) = 0.37-0.45), which according to conventional interpretation is “moderate.” For the ED visit measure, on average, the agreement between the two hospital visit rates for each hospital was 0.270 (95% confidence interval (CI) = 0.22-0.33), which according to conventional interpretation is “moderate.”

In addition, we found to achieve a reliability (ICC) of 0.4, we only require 25 patients for the inpatient admissions rate and 20 patients for the ED visit rate per performance period.

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 results of measure score reliability tests show that the data reliably distinguish performance between hospitals for both the inpatient admission and ED visit performance rates. We recommend a measurement period that is long enough to accumulate a sufficient number of patients per hospital to provide for improved reliability. This recommendation will be tested and implemented in a full national sample of hospitals when the measure goes through dry run (confidential reporting).

Here we provide you additional context for interrupting these results. This reliability result is similar to the reliability score of other NQF-endorsed outcome measures. Please note that these results are not directly comparable to the result calculated for this measure because a different measure of reliability was used in the presentation of this measure than for the measures discussed below, but this does provide an idea of the reliability scores for across endorsed, hospital-level outcome measures. For example “Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following acute myocardial infarction (AMI) hospitalization” was endorsed with a reliability score of 0.380; “Hospital 30-Day Risk-Standardized Readmission Rates following Percutaneous Coronary Intervention (PCI)” was endorsed with a reliability score of 0.3711; and “Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy” was endorsed with a reliability score of 0.335.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

We demonstrated measure validity through relevant prior validity testing we conducted for other claims-based measures, through use of established measure development guidelines and through assessment by external groups including our technical expert panel (TEP) of national experts and stakeholder organizations (face validity).

Validity of Claims-Based Measures

A number of other National Quality Forum (NQF)-endorsed measures have assessed the validity of using claims data for risk adjustment in lieu of medical record data in estimating hospital-level measure scores. CMS has validated six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission measures) with models that used medical record-abstracted data for risk adjustment. CMS has reported these findings in the peer-reviewed literature [1-6].These findings support the use of the claims-based models for public reporting.

Validity Indicated by Established Measure Development Guidelines

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures [7], CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” [8].

Validity as Assessed by External Groups (Face Validity)

Throughout measure development, we obtained expert and stakeholder input through regular discussions with our TEP and discussions with a Cancer Workgroup consisting of leaders from Prospective Payment System (PPS)-exempt cancer hospitals (Cancer Workgroup), topic-specific meetings with the seven test sites, and a public comment periods (see below for more information on each source).

- In alignment with the CMS Measures Management System, we convened a TEP to provide input and feedback during measure development. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including clinicians, patients, and individuals with experience in quality improvement, performance measurement, and healthcare disparities. The TEP had 12 members, including representation by physicians, nurses and patient advocates (see Measure Submission Form, Section Ad.1. for full membership list). We held thirteen structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members. The TEP’s role was to provide advice and feedback through all phases of the measure development process, including review and comment on evidence provided in an environmental scan, input to and reviews of measure specifications, and review and guidance relating to public comment on and testing of the measure.
- The Cancer Workgroup consisted of representatives from each of the 11 PPS-exempt cancer hospitals (see Measure Submission Form, Section Ad.1. for full membership list). The purpose of the work group was to understand current

quality measurement and improvement activities taking place at the PPS-exempt cancer hospitals and to learn their perspectives on the importance and usefulness of the measure under development.
- During development, we solicited public comment on the measure through the CMS site: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/CallforPublicComment.html. The measure specifications were posted for 45 calendar days to allow time for interested stakeholders to review and comment. 13 measure-specific comments were received, including comments from the American Hospital Association and the Alliance of Dedicated Cancer Centers. In addition, in December 2015 and January 2016 as part of the NQF Measure Applications Partnership process, the measure underwent a second public comment period. Throughout the MAP process stakeholders submitted a total of 11 unique comments.

We made modifications to the measure specifications during development (e.g., exclusion criteria and outcome definition) based on feedback. These external experts were also instrumental in highlighting potential threats to the face validity of the measure that the team investigated further to ensure measure validity—(1) attribution, (2) outcome definition, (3) exclusion criteria, and (4) risk adjustment variables. Each of these is described below.

- *Attribution*. This measure uses administration of outpatient chemotherapy to attribute a patient to a hospital, under the assumption that the hospitals administering treatment are responsible for managing the patient's clinical care and treatment-related complications. We used 2012-2013 Full Sample claims data to analyze the extent to which patients receive their outpatient care at one hospital rather than across multiple hospitals.

- *Outcome definition.* The measure assesses admissions to any acute-care hospital with one of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of an outpatient chemotherapy administration. To assess the validity of the capture of these reasons for admission, we tabulated the frequency of each diagnosis and compared it with the expectations of clinical and coding experts. In addition, to validate our approach to identifying outcomes in the claims data, we also tested the frequency of these diagnoses when: 1) listed as principal diagnosis only; 2) listed as principal or secondary diagnosis; 3) listed as principal diagnosis OR as secondary diagnosis with a principal diagnosis of cancer, to address potential differences in coding practices across hospitals.

The team tested different exclusion criteria; the results of the exclusion analyses are presented below in Section 2b.3 Exclusions Analysis. We also developed and tested a risk-adjustment methodology; the results of these analyses are presented in Section 2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures.

Process Used to Identify International Classification of Diseases, Tenth Revision (ICD-10) Codes

This application includes ICD-10 codes that correspond to all International Classification of Diseases, Ninth Revision (ICD-9) codes included in the specifications (i.e., ICD-9 to ICD-10 crosswalk). The goal was to convert this measure into a new code set, fully consistent with the intent of the original measure. For each code set containing an ICD-9-CM or ICD-9 PCS code, we identified ICD-10-CM and ICD-10 PCS codes using the General Equivalence Mapping (GEM) files made available by CMS. ICD-10 codes were searched separately to ensure capture of all relevant ICD-10-CM and PCS codes. All code sets in the attached Data Dictionary include both ICD-9 and ICD-10 codes, as appropriate.

Terminologists, clinicians, and measure development experts at the National Committee for Quality Assurance (NCQA) led and reviewed this work. All developed code sets and the ICD-9 to ICD-10 crosswalk were reviewed by NCQA's internal value set review board. The ICD-9 to ICD-10 crosswalk was also reviewed independently by a terminologist at Mathematica Policy Research. We plan to continue to evaluate the crosswalk prior to implementation.

Citations

1. Krumholz HM, Wang Y, Mattera JA, Wang Y-F, Han LF, Ingber MJ, Roman S, Normand SL. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation. 2006 Apr 4;113(13):1683-92.

2. Krumholz HM, Lin Z, Drye EE, Desai MM, Han LF, Rapp MT, Mattera JA, Normand SL. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation: Cardiovascular Quality and Outcomes. 2011 Mar 1;4(2):243-52.

3. Krumholz HM, Wang Y, Mattera JA, Wang Y-F, Han LF, Ingber MJ, Roman S, Normand SL. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation. 2006 Apr 4;113(13):1693-701.

4. Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, Schuur JD, Stauffer BD, Bernheim SM, Epstein AJ, Wang Y-F, Herrin J, Chen J, Federer JJ, Mattera JA, Wang Y, Krumholz HM. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation: Cardiovascular Quality and Outcomes. 2008 Sep;1(1):29-37.

5. Bratzler DW, Normand SL, Wang Y, O'Donnell WJ, Metersky M, Han LF, Rapp MT, Krumholz HM. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. Public Library of Science One. 2011 Apr 12;6(4):e17401.

6. Lindenauer PK, Normand SL, Drye EE, Lin Z, Goodrich K, Desai MM, Bratzler DW, O'Donnell WJ, Metersky ML, Krumholz HM. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. Journal of Hospital Medicine. 2011 Mar;6(3):142-50.

7. National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.

8. Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. *Circulation.* 2006;113(3):456-462.

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

Validity as Assessed by External Groups

Several analyses were conducted to address potential threats identified to measure validity (described in Section 2b2.2). The results of the analyses on the identified potential threats to validity are provided below. Following these additional analyses, feedback from stakeholders, including all members of the TEP, affirmed that the measure as specified can be used to distinguish between better and worse quality hospitals.

Attribution. The measure uses an administration of outpatient chemotherapy to attribute a patient to that hospital, under the assumption that hospitals administering treatment are responsible for managing the patient's complications. Using the 2012-2013 Full Sample claims data, testing showed that only 5% of cancer patients in our cohort received an outpatient chemotherapy treatment from more than one hospital. That is, results from testing showed that patients typically receive their treatment at one hospital, which supports our methodology for attributing the management of outpatient care to the reporting hospitals. The TEP reviewed these findings and agreed with our methodology.

Outcome definition. The list of diagnoses includes those that are common causes of potentially avoidable hospital admissions and ED visits among cancer patients receiving outpatient chemotherapy and that can be improved by timely provision of evidence-based care in the outpatient or ambulatory care setting. Through feedback from stakeholders and

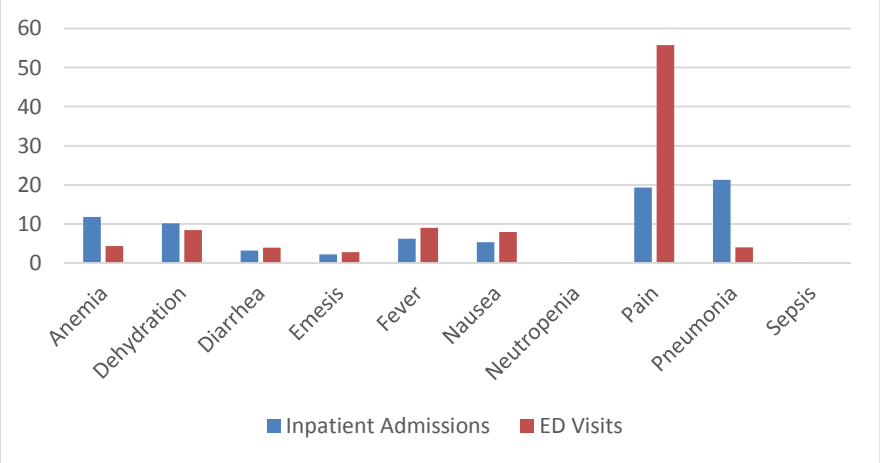
experts and our testing results, we made modifications to the specification to better capture the intended population. These changes included:

- As a single billing code for neutropenic fever is not available in ICD-9, we separated this diagnosis into neutropenia and fever.
- Neutropenia is rarely listed on a claim because the primary reason for admission is typically an infection and not neutropenia. Based on clinical input from our TEP, we then expanded the list of diagnoses to also include pneumonia and sepsis.
- To enhance the clinical face validity of the measure and ensure that hospital billing practices would not impact performance, we expanded the measure algorithm to look for inpatient and ED claims within 30 days of an outpatient chemotherapy treatment where the principal diagnosis on the claim is either: (a) one of the ten qualifying diagnoses; or (b) any cancer diagnosis, accompanied by one of the ten qualifying diagnoses listed as a secondary diagnosis on the same claim.

Testing results revealed that the most common reasons for admission included pneumonia (21 percent of admissions), pain (19 percent), and anemia (12 percent); the most common reasons for ED visits included pain (56 percent of ED visits), fever (9 percent), and dehydration (8 percent). Additionally, very few patients (<1%) were admitted or seen at an ED with a principal diagnosis of neutropenia or sepsis. The TEP and billing experts at our test sites affirmed the finding that few claims listed neutropenia as the principal diagnosis because the primary reason for an admission or ED visit would usually be infection and not neutropenia. To capture these patients, the outcome also includes patients with a principal diagnosis of pneumonia and sepsis. Furthermore, the TEP supported continued inclusion of all ten outcome-qualifying diagnoses, as even those with low frequency are potentially avoidable.

The team also held discussions with coding experts at one of our test sites to better understand whether coding practices are a potential threat to the validity of including these ten principal diagnoses in our outcome definition. When asked to select the top three diagnoses from our list, they chose pain, fever, and dehydration, which align with our findings. The coding experts confirmed that pain is a frequent occurrence and that they would expect a majority of admissions to be related to pain. Coding experts also agreed the low numbers of patients with neutropenia may be a result of a patient being admitted before lab results are returned, and stated their providers may not be documenting neutropenia very well as the principal or secondary diagnosis. In discussing the measure more broadly, the coding experts did not have any concerns with the accuracy of coding for these conditions as the principal diagnosis, when appropriate, and supported our limitation of the diagnoses to principal or secondary with a principal diagnosis of cancer, given the intent of the measure.

Percent of Patients by Outcome Qualifying Diagnosis



We also assessed the algorithm used to identify outcomes through empirical analyses and feedback from our TEP. Based on our desire to make the measure as broad and meaningful as possible, we finalized an algorithm that counts stays/ED visits with one of the ten qualifying diagnoses in the principal diagnosis position, *or* with one of the ten diagnosis in a

secondary position if accompanied by a principal diagnosis of cancer; this approach addresses potential differences in coding practices across hospitals which are missed when we consider only those in the principal position. The observed rates of outcomes with this broader definition are slightly higher than the observed rates when we limited to principal diagnosis only: 10.3 percent versus 7.2 percent for inpatient stays, and 4.2 percent versus 3.9 percent for ED visits.

Process Used to Identify International Classification of Diseases, Tenth Revision (ICD-10) Codes
We did not perform ICD-10 testing because no ICD-10 data were available. We plan to continue to evaluate the ICD-10 codes prior to implementation.

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

Qualitative and quantitative validity testing results demonstrate TEP and external stakeholder agreement with the overall face validity of the measure as specified.

Analysis of claims showed that the attribution methodology is operating as intended. The ten diagnoses specified in the measure outcome are valid, as confirmed through empirical analysis and expert feedback. In addition, the algorithm used to define the outcome was shown to be valid and aligned with the intent of the measure. Furthermore (presented below in Section 2b3) experts agreed that the measure exclusions are appropriate and the risk factors included in the risk-adjustment model are variables the TEP expect to affect outcomes, which further enhances face validity. For further discussion on exclusions and risk adjustment, please see Section 2b.3 Exclusions Analysis and Section 2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures.

_____

2b3. EXCLUSIONS ANALYSIS

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

We evaluated four possible exclusions. First, we considered the removal of patients with leukemia because of the high toxicity of treatment and expected readmissions due to relapse. Second, we considered the removal of patients who do not have a full year of prior enrollment data to ensure complete data for the risk-adjustment model. Third, we considered removal of patients who do not have at least one chemotherapy treatment followed by 30 days of enrollment for full data availability to identify outcomes. Lastly, we considered the removal of patients younger than 65 years of age because patients aged 18-64 enrolled in Medicare may be systematically different than those patients 65 and older.

We reviewed each of these exclusions with our TEP. Expert input raised concerns about the exclusion of patients aged 18-64 expressing a desire for a broad cohort and no clinical reason to exclude this group; we therefore conducted additional data analyses. We explored the appropriateness of including these patients by (1) reviewing patient characteristics separately for these two subsets, (2) reviewing the observed performance rates for the two separate subsets, and (3) fitting the risk-adjustment model separately for these two subsets.

Expert input determined the first three exclusions to be clinically relevant or required for data completeness (see Measure Submission Form, Section S.10 for more information). For the first three exclusions, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion. We then looked at the distribution of the exclusions across hospitals. Lastly, we calculated the observed performance rate with and without accounting for exclusions. The results are presented below.

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

We explored the potential bias of including patients aged 18-64 from the cohort by (1) reviewing patient characteristics separately for these two subsets, (2) reviewing the observed performance rates for the two separate subsets, and (3) fitting the risk-adjustment model separately for these two subsets. We found that patients aged 18-64 represent 13% of the final measure cohort, and while the younger population has higher observed outcome rates, the risk-adjustment model parameter estimates were similar for both subsets of patients. Based on these findings, as well as the recommendation of our TEP, we determined there was not a strong statistical or clinical reason to exclude the younger patients from the measure cohort; all adult patients 18 years and older remain in the eligible cohort.

Then, applying our inclusion criteria (Medicare FFS patients aged 18 years and older with a cancer diagnosis who received chemotherapy treatment in a hospital outpatient department during the performance period), 320,516 unique patients were included in the initial measure cohort. We then applied the follow exclusion criteria (note that these groups are not mutually exclusive; see the Measure Submission Form, Sections S.10 and S.11, for full list of exclusions and codes):

1) Patients with a diagnosis of leukemia at any time during the performance period (25,714 patients; 8% of initial measure cohort)
2) Patients who were not enrolled in Medicare FFS Parts A and B in the year prior to the first outpatient chemotherapy treatment during the performance period (55,926 patients; 17% of initial measure cohort)
3) Patients who did not have at least one chemotherapy treatment with enrollment in Medicare FFS Parts A and B in the 30-days after the procedure (18,193 patients; 6% of initial measure cohort)

A total of 80,070 unique patients, or about 25% of the eligible cohort, were excluded due to these exclusions.

Table 1 provides the distribution of the percentage of patients excluded for each exclusion across hospitals. This table indicates that there is modest variation in the number of cases excluded within hospitals.

Table 1. Distribution of Percentage of Patients Excluded across Hospitals (N=3,765 hospitals)

Exclusion	25th percentile	50th percentile	75th percentile
Diagnosis of leukemia	0	5.1	9.5
12 month prior enrollment	0	12.5	20.7
30-day continued enrollment	0	1.5	7.0

Using the eligible cohort prior to exclusions, we calculated the following observed rates: 14.1 percent of patients receiving treatment in a hospital outpatient setting were admitted within 30 days of chemotherapy, 5.2 percent visited the ED, such that nearly 20 percent of the initial cohort experienced an outcome. Using the cohort after applying exclusion criteria, we calculated the following observed rates: 10.3 percent of patients receiving treatment in a hospital outpatient setting were admitted within 30 days of chemotherapy and 4.2 percent visited the ED, such that just under 15 percent of the final cohort experiencing an outcome. As expected, primarily due to the higher admission rate expected for leukemia patients, the observed rates prior to application of exclusion criteria were higher.

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

After extensive literature review, examination of existing measures, consideration of requirements to have adequate risk adjustment and identification of admissions or ED visits, and discussion with the working group and TEP members,

we determined the three exclusion criteria are necessary for a valid measure. The goal was to be as inclusive as possible while creating a clinically coherent cohort. In other words, the measure population had to be sufficiently similar in terms of the procedure and outcome risk profile. This was to ensure risk adjustment can be performed adequately.

Testing of the distribution of exclusion criteria across hospitals suggests modest variation among providers. The uneven distribution of excluded populations and procedures supports our decision that these exclusions are required. Failure to exclude these populations may distort the measure score and unfairly disadvantage certain hospitals. Additional rationales for exclusions are detailed in the Measure Submission Form, Section S.10. After exclusions were applied, the measure captured the majority (75.0%) of all qualifying patients. The exclusions are very narrowly targeted and necessary to ensure a clinically coherent measure cohort and a cohort with complete data available for risk adjustment and identification of admissions or ED visits.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☒ **Statistical risk model with** 20 **risk factors for the inpatient admission outcome model and 15 risk factors for the ED visit outcome**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

Not applicable. This measure is risk-adjusted.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Our approach to risk adjustment is tailored to, and appropriate for, a publicly reported outcome measure as articulated in published scientific guidelines [6,8].

The measure has two mutually exclusive outcomes: (1) patients in the cohort admitted to any acute-care hospital with one of the following diagnoses—anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia, or sepsis—within 30 days of an outpatient chemotherapy administration at the reporting hospital, and (2) patients in the cohort that did not have a qualifying inpatient admission, but were seen at any ED with one of the qualifying diagnoses within 30 days of an outpatient chemotherapy administration at the reporting hospital. As a result, we developed two risk-adjustment models, one for each dependent variable—inpatient admissions and ED visits. The same variable selection process was used for both models. We provide the sequential method in selecting patient factors below.

Candidate Risk-Adjustment Variables
Candidate risk-adjustment variables were patient-level risk adjustors that are expected to be predictive of the outcomes based on prior literature, clinical judgment, and empirical analysis. We limited our initial selection of candidate variables for inclusion in our preliminary risk-adjustment model to variables with a strong clinical rationale for inclusion as identified in the literature and through clinical expert input. Identification of these variables is described below. See attached Data Dictionary, sheet "2b4.3 Candidate Variables" for the list of candidate variables.

Demographic variables: In alignment with the specifications of other NQF endorsed claims-based outcome measures, as well as the NQF guidelines at the time of development, we included age and sex as candidate covariates. [Note: Due to changes in NQF policy, additional socio-demographic factors were considered during continued assessment of this measure as described later in this section.]

Comorbidities: The model adjusts for case mix differences based on the comorbidities of the patient at the time of the first outpatient chemotherapy treatment during the performance period. We define comorbidities using Condition Categories (CCs) from Version 12 (V12) of the CMS-HCC risk-adjustment model, which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes. With a subset of our TEP, we reviewed all 189 CCs to determine the clinical appropriateness and prevalence within the cohort) for potential inclusion in the model.

Specific considerations included the number of patients in our cohort potentially affected, whether the condition affects admission for one of the ten outcome-qualifying diagnoses, and whether inclusion of the condition in the model would incentivize appropriate treatment, even when that variable is theoretically unrelated to admission for one of the identified reasons. For example, patients with diabetes may have gastric paresis, a condition that slows emptying of the stomach and increases the likelihood of nausea. Adjusting for diabetes might reduce incentives to provide chemotherapy drugs that would be less likely to cause nausea in patients with diabetes and gastric paresis. The CCs selected for inclusion were bundled with other clinically related CCs for empirical assessment of significance within the model. The result was nine bundled CCs—diabetes, metabolic disorders, gastrointestinal (GI) disorders, psychiatric disorders, neurological conditions, cardiovascular disease, respiratory disorders, renal disease, and other injuries.

Indicator of disease severity: We explored cancer type as an indicator of disease severity available in claims data by assessing the distribution of patients across a granular level of cancer diagnoses. In conjunction with a subset of our TEP, we aggregated these granular cancer types into nine clinically related and decently sized groupings—breast cancer, digestive cancer, genitourinary cancer, respiratory cancer, lymphoma, prostate cancer, secondary cancer of the lymph nodes, secondary cancer of solid tumor, and other cancers.

Exposure: We also assessed the number of chemotherapy treatments during the performance period (that is, exposure). The exposure variable is necessary because the measure estimates the risk-adjustment models at the patient level and the number of outpatient chemotherapy treatments varies by patient. Patients with more treatments during the period have an increased probability of experiencing an outcome because the algorithm looks for an outcome after each treatment. The exposure variable is the count of outpatient chemotherapy administrations the patient experienced at the attributed hospital during the performance period.

Interactions: Through discussion with our Expert Working Group (see Measure Submission Form, Section Ad.1. for full membership list), we determined the most clinically relevant interactions are likely to be between the age variable and the different cancer types. Based on this input, we tested age-cancer type interaction terms as candidate covariates.

<u>Variable Selection</u>

To select the final variables to include in the risk-adjustment model, we fitted a logistic regression model to predict the outcome with the candidate variable set. To develop a parsimonious model, we then removed non-significant variables from the initial model using a stepwise purposeful selection method described by Hosmer and Lemeshow [1,3]. Our goal was to minimize the number of variables in the model while preserving model performance (as measured by the c-statistic). During this process, for each of the two models, the least significant variable in the model was removed one at a time until only statistically significant ($p<0.05$, assessed using a likelihood ratio test) variables remained in the model. Interaction terms between age and cancer type were tested and were only retained in the model if significant at a level of $p<0.01$. The higher threshold for statistical significance of interaction terms was to ensure that only interactions that have a higher likelihood of being true interactions were included.

The attached Data Dictionary, sheet "S.15 Risk Model Specs" indicates the final risk variables selected, the codes used to define the risk variables for our statistical model, and their odds ratios and 95% CIs.

SDS Conceptual Model

Following the selection process for clinically-relevant risk factors described above, we undertook an assessment on the need to incorporate additional SDS factors into our risk-adjustment model. In this section, we describe the conceptual model that guided our work. In Section 1.8, we described the three variables used in our analysis (race, Medicaid dual eligible status, and neighborhood SES factors composited into the AHRQ SES Composite Index).

The potential causal pathways by which SDS risk factors influence the risk of admission or ED visit following outpatient chemotherapy are varied and complex. The presence of disparities in chemotherapy outcomes are due to multiple complementary causes. To help inform our conceptualization of the pathways by which SDS factors affect admissions and ED visits for patient receiving chemotherapy treatment in a hospital outpatient setting, we performed a literature search. The studies indicated that individuals that identify as a racial minority, with low socioeconomic status (SES), with charity care or self-pay insurance, are women, or are unmarried were more likely to experience a gap in cancer care in the outpatient chemotherapy setting than their counterparts. Please refer to question 1 of the "ChemoMeasure_NQF Appendix_SDS" for more information on the literature review.

The following highlights possible SDS-related conceptual pathways that are important to consider:

1. **Relationship of SDS with health.** Those that face sociodemographic disadvantages usually have worse health status, which in run leads to worse health outcomes, than those that do not experience these disadvantages. This means that chemotherapy patients that have lower health literacy, income, education, and no insurance might experience a higher symptom burden or have greater disease severity, and in turn have more ED visits and hospital admissions due to having worse health status in general. This pathway could be accounted for within the existing clinical risk-adjustment variables in the current model.

2. **Access to care.** Limited access to health care may prevent individuals from early detection of cancer, making them more likely to be diagnosed with late-stage cancer that could have been treated more effectively or cured if diagnosed earlier [7]. Worse access to care also impacts patients ability to contact their physicians when they are experiencing cancer-related symptoms or adverse effects from treatment, which may make them more likely to experience ED visits, hospital admissions, ambulance use, and hospital mortalities compared to cancer patients that are diagnoses at an earlier stage [5].

3. **Differential care across hospitals.** Cancer patients at minority-serving hospitals are less likely to receive adequate pain treatment [2]. Poor and minority patients are also more likely to be seen in safety-net hospitals and these hospitals may lack the financial resources to make certain services available, such as specialized palliative care teams, making these patients more likely to require acute care, such as an ED visit or hospital admission, for symptom management.

The combination of treatment disparities, increase symptom occurrence and severity, and inadequate pain management may place minority cancer patients at greater risk of experiencing a gap in outpatient chemotherapy care, which may increase the likelihood of ED visits and hospital admissions.

Citations

1. "Disparities in Cancer Care." Journal of Oncology Practice / American Society of Clinical Oncology, vol. 2, no. 5, 2006, pp. 234-239.

2. Fisch, M.J., J.W. Lee, M. Weiss, L.I. Wagner, V.T. Chang, D. Cella, J.B. Manola, L.M. Minasian, W. McCaskill-Stevens, T.R. Mendoza, and C.S. Cleeland. "Prospective, Observational Study of Pain and Analgesic Prescribing in Medical Oncology Outpatients with Breast, Colorectal, Lung, or Prostate Cancer." Journal of Clinical Oncology, vol. 30, no. 16, 2012, pp. 1980–1988.

3. Hosmer DW LS. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* New York: Wiley; 1999.

4. Killelea, B. K., V. Q. Yang, S. Y. Wang, B. Hayse, S. Mougalian, N. R. Horowitz, A. B. Chagpar, L. Pusztai, and D. R. Lannin. "Racial Differences in the use and Outcome of Neoadjuvant Chemotherapy for Breast Cancer: Results from the National Cancer Data Base." Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, vol. 33, no. 36, 2015, pp. 4267-4276.

5. Kotajima, F., K. Kobayashi, H. Sakaguchi, and M. Nemoto. "Lung Cancer Patients Frequently Visit the Emergency Room for Cancer-Related and -Unrelated Issues." Molecular and Clinical Oncology, vol. 2, no. 2, 2014, pp. 322-326.

6. Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. *Circulation*. 113 (3): 456-462.

7. National Cancer Institute. "Cancer Health Disparities." Available at: [http://www.cancer.gov/about-nci/organization/crchd/cancer-health-disparities-fact-sheet]. Accessed on January 19, 2016.

8. Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Stat Sci.* 2007; 22 (2): 206-226
Hosmer DW LS. *Applied Logistic Regression.* New York: Wiley; 2000.

9. Shavers, V. L. and M. L. Brown. "Racial and Ethnic Disparities in the Receipt of Cancer Treatment." Journal of the National Cancer Institute, vol. 94, no. 5, 2002, pp. 334-357.

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

The final models retained 20 variables in the inpatient admission outcome model and 15 variables in the ED outcome model with p values <0.05. The (age x cancer type) interaction terms were retained if p for interaction was <0.01. For the inpatient admission outcome model, only the interaction of (age x digestive cancer) was significant (p-value for interaction <0.001). However, due to the minimal improvement in model fit (AIC (76245 → 76238) and c-statistic (0.725 → 0.725) and our desire to create the most parsimonious model, we did not include any interaction terms in our final model. No interaction terms met this criterion for the ED visit outcome model.

In addition, the final model did not include SDS variables. See Section 2b4.4b for more information.

The following variables were selected as the final risk-adjustment variables for the inpatient admission outcome model. In addition, Table 3 of the attached Measure technical Report includes the coefficients, odds ratios, and confidence intervals for each variable.

- Age (continuous)
- Sex (male)
- Exposure (Number of hospital OPD chemotherapy treatments during period)
- Respiratory disorders (CC 107 – 110)
- Renal disease (CC 128 – 131)
- Diabetes (CC 15 – 20)
- Other injuries (CC 162)
- Metabolic disorder (CC 21-24)
- Gastrointestinal disorder (CC 25-36)
- Psychiatric disorder (CC 48-66)
- Neurological conditions (CC 67-76)
- Cardiovascular disease (CC 77-106)
- Breast cancer (ICD codes 174.0-175.9)
- Digestive cancer (ICD codes 150.0–159.9)

Respiratory cancer (ICD codes 160.0–165.9)
Lymphoma (ICD codes 200.00–203.82)
Other cancer (ICD codes 140.0–149.9, 170.0–173.99, 176.0–176.9, 179, 182.0–182.8, 190.0–199.2, 209.00-209.36)
Prostate cancer (ICD codes 185)
Secondary – lymph (ICD codes 196.0-196.9)
Secondary – solid (ICD codes 197.0-198.82, 209.70-209.79)

The following variables were selected as the final risk adjustment variables for the ED visit outcome model. In addition, Table 4 of the attached Measure technical Report includes the coefficients, odds ratios, and confidence intervals for each variable.

Age (continuous)
Sex (male)
Exposure (Number of hospital OPD chemotherapy treatments during period)
Respiratory disorders (CC 107 – 110)
Other injuries (CC 162)
Gastrointestinal disorder (CC 25-36)
Psychiatric disorder (CC 48-66)
Neurological conditions (CC 67-76)
Cardiovascular disease (CC 77-106)
Breast cancer (ICD codes 174.0-175.9)
Digestive cancer (ICD codes 150.0–159.9)
Respiratory cancer (ICD codes 160.0–165.9)
Lymphoma (ICD codes 200.00–203.82)
Other cancer (ICD codes 140.0–149.9, 170.0–173.99, 176.0–176.9, 179, 182.0–182.8, 190.0–199.2, 209.00-209.36)
Prostate cancer (ICD codes 185)
Secondary – lymph (ICD codes 196.0-196.9)
Secondary – solid (ICD codes 197.0-198.82, 209.70-209.79)

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

As described in Section 1.8, we used three variables in our SDS analysis—(1) race (black, other), (2) Medicaid dual eligible status, and (3) neighborhood SES factors composited into the AHRQ SES Composite Index. We conducted several analyses, presented below, including (1) variation in patient SDS risk factors across hospitals; (2) the association between SDS variables and the outcomes; (3) the impact of including SDS variables as part of risk-adjustment on model performance; and (4) the impact of including SDS variables as part of risk-adjustment on hospital rankings. Key findings and our conclusion are described below. A complete summary of our SDS assessment, including analysis tables, is provided in the separate appendix titled "ChemoMeasure_NQF Appendix_SDS."

<u>Variation in Prevalence of SDS Risk Factors across Hospitals</u>

There is substantial variation in the prevalence of black, Medicaid dual-eligible, and AHRQ SES Composite Index across patients in the measure cohort across hospitals. For the measure, the percentage of patients who are black ranges from 0% to 100% across hospitals, with a median of 0.7% (interquartile range [IQR] 0%-10.6%). The percentage of patients who are Medicaid dual eligible ranges from 0% to 100% across hospitals, with a median of 18.1% (IQR 9.0% - 30.7%). The percentage of patients with low AHRQ SES Composite Index ranges from 0% to 100% across hospitals, with a median of 19.0% (IQR 2.2% - 52.5%).

Association between SDS variables and the outcomes

The patient-level relationship was described in the NQF Submission Form, Section 1b.4. On the *patient-level*, our analysis shows that "low SDS" patients (as characterized by three individual indicators: Medicaid dual-eligibility, race as black, and AHRQ SES Composite Index) receiving hospital-based outpatient chemotherapy are more likely to have an inpatient admission and emergency department (ED) visit within 30 days than "non-low SDS" patients.

On the *hospital-level*, no between-hospital effects were observed for hospital case-mix by Medicaid dual-eligibility, race, or the AHRQ SES Composite Index. Specifically, there was no clear relationship between the median risk-standardized rates and hospitals' case mix by these three SDS factors. In addition, the distributions of risk-standardized rates overlapped significantly across hospitals grouping by these three SDS factors, suggesting that hospitals caring for a greater percentage of low SDS patients have similar rates of inpatient admission and ED visits within 30 days of hospital-based outpatient chemotherapy. For example, the hospitals in the lowest quartile of proportion of black patients had a median risk-adjusted admission rate of 10.2, the second quartile had a rate of 10.6, third quartile had a median rate of 10.1, and the top quartile of hospitals with proportion of black patients had a rate of 10.2. For full presentation of results see Table 5 in the separate appendix titled "ChemoMeasure_NQF Appendix_SDS."

Risk-adjustment model diagnosis and performance with and without SDS variables

There is no multicollinearity between the SDS variables and the existing covariates. Also, there is no multicollinearity within SDS variables. Variance inflation factors (VIF) for all covariates are close to 1 (max VIF = 1.57 for IP model and 1.34 for ED model).

All three SDS variables exhibit statistically significant association with the outcome, and their directions make sense in explaining their associations with the outcome.

Models exhibit similar performance with and without including SDS variables in the risk adjustment. Specifically,

- C-statistics exhibit very similar model discrimination between risk adjustment using original risk factors and using original risk factors plus SDS variables. For example, for the Validation Split Sample, the inpatient admission measure C-statistics are 0.725 for the model that does not adjust for SDS variables and 0.728 for the model that adjusts for SDS variables. For the ED visit measure, the C-statistics are 0.636 without adjusting for SDS and 0.644 when adjusting for SDS.
- The model calibration results are very similar between risk adjustment using original risk factors and using original risk factors plus SDS variables.
- The results of overfitting indices remained similar with and without adding SDS variables in the risk-adjustment model.

Hospital rankings after considering SDS variables

There is very high agreement of hospital rankings between risk-adjustment models which incorporate SDS variables and those that do not (Spearman rank correlation = 0.988 for the inpatient admission model and 0.984 for the ED visit model), suggesting that accounting for the SDS factors will not have a major impact on hospital rankings.

There are clear patient-level effects, but at the hospital level, accounting for patient SDS factors has minimal to no impact on model performance or hospital performance ranking for both the admission or ED measure, indicating that the added risk of being low SDS is captured within current risk variables and arguing against inclusion of patient SDS factors in the chemotherapy measure. Given these findings, we did not include SDS factors in the risk-adjustment model for this measure.

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

We assessed adequacy of the patient-level risk-adjustment models (described above). We evaluated the model performance first in the 2012-2013 Medicare FFS Development Split Sample. We then re-tested the model performance in the 2012-2013 Medicare FFS Validation Split Sample. We did this separately for both the inpatient admission outcome model and the ED visit outcome model.

Using the 2012-2013 Medicare FFS Development Split Sample, we computed three summary statistics for assessing the risk-adjustment model performance: area under the receiver operating characteristic (ROC) curve (c-statistic), predictive ability, and over-fitting indices. We then compared the model performance in the development sample with its performance in the validation sample.
The c-statistic is a measure of how accurately a statistical model is able to distinguish between a patient with and without a hospital visit. For binary outcomes, the c-statistic is identical to the ROC. A c-statistic of 0.50 indicates random prediction, implying that patient risk factors contribute no additional information. A c-statistic of 1.0 indicates perfect prediction, implying that patients' outcomes can be predicted completely by their risk factors.

Discrimination in predictive ability measures the ability to distinguish high-risk from low-risk subjects. Good model discrimination is indicated by a wide range between the lowest and highest deciles.

We assess model calibration by calculating over-fitting indices. Over-fitting refers to the phenomenon in which a model describes the relationship between predictive variables and outcome well in one group of patients, but fails to provide valid predictions in another distinct group of patients. Over-Fitting indices (γ0, γ1) provide evidence of over-fitting and require several steps to calculate. The mathematical process is described here: Let b denote the estimated vector of regression coefficients. Predicted Probabilities (p) = 1/(1+exp{-Xb}), and Z = Xb (e.g., the linear predictor that is a scalar value for everyone). A new logistic regression model that includes only an intercept and a slope by regressing the logits on Z is fitted in the validation sample; e.g., Logit(P(Y=1|Z)) = γ0 + γ1Z. Estimated values of γ0 far from 0 and estimated values of γ1 far from 1 provide evidence of over-fitting.

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

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

Inpatient admission outcome model
2012-2013 Medicare FFS Development Split Sample results:
c-statistic=0.73
Predictive ability (lowest decile %, highest decile %): 2.09-27.70%

2012-2013 Medicare FFS Validation Split Sample results:
c-statistic=0.72
Predictive ability (lowest decile %, highest decile %): 2.16-27.98%

ED visit outcome model
2012-2013 Medicare FFS Development Split Sample results:
c-statistic=0.63
Predictive ability (lowest decile %, highest decile %): 1.91-8.33%

2012-2013 Medicare FFS Validation Split Sample results:
c-statistic=0.64
Predictive ability (lowest decile %, highest decile %): 1.93-8.22%

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Inpatient admission outcome model
2012-2013 Medicare FFS Development Split Sample results:
Calibration: (γ0, γ1) = (0,1)

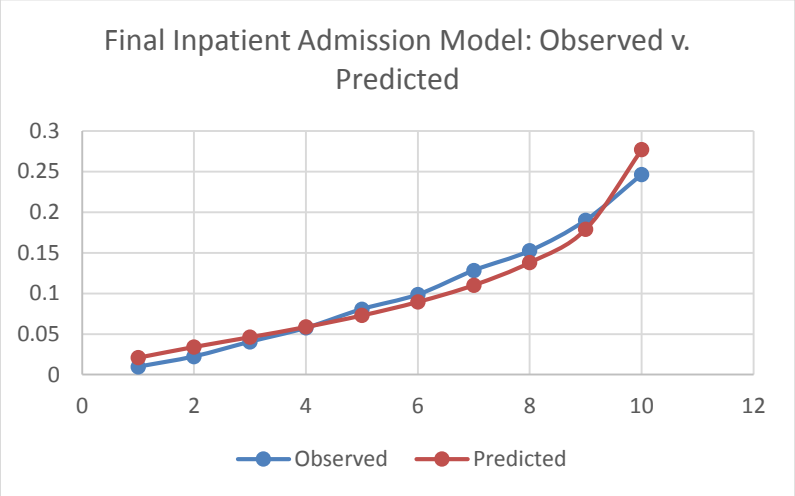
2012-2013 Medicare FFS Validation Split Sample results:
Calibration: (γ0, γ1) = (0.01, 1.00)

ED visit outcome model
2012-2013 Medicare FFS Development Split Sample results:
Calibration: (γ0, γ1) = (0,1)

2012-2013 Medicare FFS Validation Split Sample results:
Calibration: (γ0, γ1) = (-0.04, 0.99)

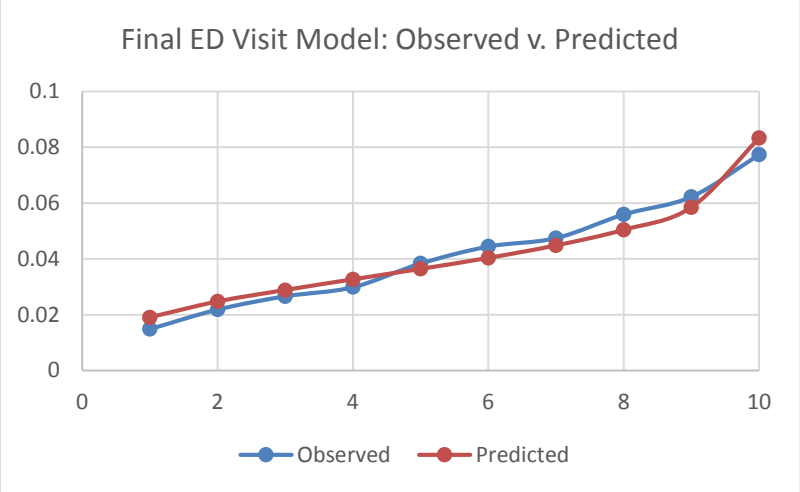
2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Inpatient admission outcome model: Plot of observed vs. predicted values for risk deciles (2012-2013 Medicare FFS Development Split Sample)



A second plot using 2012-2013 Medicare FFS Validation Split Sample showed very similar results.

ED visit outcome model: Plot of observed vs. predicted values for risk deciles (2012-2013 Medicare FFS Development Split Sample)



A second plot using 2012-2013 Medicare FFS Validation Split Sample showed very similar results.

2b4.9. Results of Risk Stratification Analysis:

Not applicable. This measure is not risk stratified.

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

For both models, model performance was similar in the development and validation datasets, with strong model discrimination and fit. Predictive ability was also similar across datasets. The c-statistics of 0.73 (inpatient) and 0.63 (ED visit) indicate good model discrimination. The models indicated a wide range in predictive ability between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects. The calibration value of close to 0 and close to 1 indicates good calibration of the model. Additionally, the risk decile plots show that the model performs similarly in each of the risk deciles across a broad range of risk.

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

Not applicable

_____

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

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

The measure estimates hospital-specific risk-adjusted rates of potentially preventable inpatient admissions or ED visits for cancer patients aged 18 years or older receiving chemotherapy treatment in the hospital outpatient setting using a hierarchical logistic regression model. The method for discriminating hospital-level performance for public reporting has not been determined (e.g., better or worse than national rate).

Here we review the distribution of observed and risk-adjusted rates across all hospitals to show meaningful differences in performance across hospitals.

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

Table 2 shows the distribution of both the hospital-specific observed and risk-adjusted performance rates across all hospitals. The observed rates for qualifying inpatient admissions range from 0 to 100 with an average hospital rate of 8.3. The associated distribution for risk-adjusted rates ranges from 6.0 to 24.9. For the ED visits outcome, the rates across hospitals also range from 0 to 100 with an average hospital rate of 4.3. The associated distribution for risk adjusted rates ranges from 2.1 to 7.5.

Table 2. Outcome rates, among hospitals with any case size

Result	Mean	Stnd Dev	Min	25th Pctl	Median	75th Pctl	Max
Observed Admission Rate	8.3	0.11	0.0	0.0	6.5	12.1	100.0
Risk-Adjusted Admission Rate	10.4	1.28	6.0	9.8	10.2	10.8	24.9
Observed ED Visit Rate	4.3	0.09	0.0	0.0	1.4	5.3	100.0
Risk-Adjusted ED Visit Rate	4.2	0.53	2.1	4.0	4.1	4.4	7.5

Source: Claims from Medicare Parts A and B from July 1, 2012 through June 30, 2013.

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

The results suggest that there are meaningful differences in the quality of care received for adult cancer patients receiving chemotherapy treatment in the hospital outpatient setting. Overall observed rates show that an average of 12% of the cohort experience a potentially preventable outcome, with variation in performance across hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

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

2b6.1. Describe the method of testing conducted to demonstrate comparability of 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*)

Items 2b6.1-2b6.3 skipped, as this measure has only one set of specifications.

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

Items 2b6.1-2b6.3 skipped, as this measure has only one set of specifications.

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

are the norms for the test conducted)

Items 2b6.1-2b6.3 skipped, as this measure has only one set of specifications.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Not applicable

2b7.2. 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

2b7.3. 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)*

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.**

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.**

No feasibility assessment **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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Measure development and testing showed that the measure cohort can be defined and outcomes reported using routinely collected Medicare claims data. This measure is not in operational use.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Not applicable. There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

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.

Planned	Current Use (for current use provide URL)
Public Reporting	

4a.1. **For each CURRENT use, checked above, provide:**

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

Not applicable. Measure is not yet in use.

4a.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 is not currently publicly reported or used in an accountability application because it has only recently completed development and is being submitted to NQF for initial endorsement.

4a.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.*)
The measure may ultimately be used in one or more CMS programs, such as:
-Hospital Outpatient Quality Reporting Program
-PPS-Exempt Cancer Hospital Quality Reporting Program

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Not applicable

4b.2. **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.**
Since this measure is not yet in use, there are no performance results to assess improvement.

We expect there to be improvement in measure scores over time since publicly reported measure scores can reduce adverse patient outcomes associated with poorly managed outpatient care by capturing and making more visible to providers and patients all potentially preventable hospital visits following chemotherapy treatment in the hospital outpatient setting.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We did not identify any unintended consequences during measure development or model testing. However, during the NQF Measure Applications Partnership (MAP) review of this measure in December 2015, we received concerns about a possible unintended consequence related to treatment decisions and underuse of appropriate care. The concern was that the measure might indirectly discourage more aggressive treatment plans that would have had clinical benefits. However, the purpose of the measure is to open lines of communication between the patient and provider on risks and preventative actions that can be taken for each type of treatment, and set the expectations for the patient so they can make more informed decisions on healthcare utilization as well [1]. Furthermore, the measure is risk adjusted to help account for the variation in patient mix and aggressiveness of treatment. For example, aggressiveness can range by cancer type and age, which are accounted for in our model. We also adjust for the number of treatments which may also be an indicator of aggressiveness. Lastly, the measure rate is not intended to be zero and CMS recognizes that not all admissions and ED visits are avoidable. Improving patient/provider communication and appropriately adjusting the model mitigates the risk of the unintended consequences.

We are committed to monitoring this measure's use and assessing potential unintended consequences over time.

Citation

1. Aprile, G., F.E. Pisa, A. Follador, L. Foltran, F. De Pauli, M. Mazzer, S. Lutrino, C.S. Sacco, M. Mansutti, and G. Fasola. "Unplanned Presentations of Cancer Outpatients: A Retrospective Cohort Study." Supportive Care in Cancer, vol. 21, no. 2, 2013, pp. 397–404.

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)**
0383 : Oncology: Plan of Care for Pain – Medical Oncology and Radiation Oncology (paired with 0384)
0384 : Oncology: Medical and Radiation - Pain Intensity Quantified
1628 : Patients with Advanced Cancer Screened for Pain at Outpatient Visits

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Cancer – fatigue/anemia: percentage of patients seen for an initial visit or any visit while undergoing chemotherapy at a cancer-related outpatient site for whom there was an assessment of the presence or absence of fatigue (RAND Corporation)

5a. Harmonization

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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
All four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) focus on cancer patients receiving outpatient chemotherapy; however, there are some key differences in measure scope and measure type. Measure scope: Each of the four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) narrowly focuses on pain management and/or fatigue/anemia. The proposed measure does not target a specific symptom, but rather assesses the overall management of 10 important symptoms and complications that we identified as being more frequently cited in literature as reasons for ED visits and inpatient admissions following outpatient chemotherapy. Measure type: The four related measures (NQF 0383a, NQF 0384a, NQF 1628, and Cancer – fatigue/anemia) are all process measures encouraging the use of screening and care plans to improve care. The proposed measure is an outcome measure not encouraging or measuring specific processes to detect and treat these conditions, but rather assessing the outcomes of the care being provided. The four process measures, which are not risk-adjusted, support the intent of the measure by reinforcing that those providing outpatient care should screen for and manage symptoms such as pain.

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.
Attachment **Attachment:** ChemoMeasure_NQF_Appendix_SDS_and_Technical_Report.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services (CMS)
Co.2 **Point of Contact:** Vinitha, Meyyur, Vinitha.Meyyur@cms.hhs.gov, 410-786-8819-
Co.3 **Measure Developer if different from Measure Steward:** Mathematica Policy Research
Co.4 **Point of Contact:** Christine, Holland, CHolland@mathematica-mpr.com, 202-484-5271-

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.

CORE/Mathematica convened a 12-member technical expert panel (TEP) of physicians, nurses, patient advocates, and experts in quality improvement to provide input on key methodological decisions. The TEP convened at the initiation of the project and was used throughout development with minimal changes to its composition.

TEP Members:

- Peter B. Bach, MD, MAPP—Memorial Sloan-Kettering Cancer Center, Center for Health Policy and Outcomes (Director); New York, NY
- Stephen Edge, MD—Baptist Cancer Center (Cancer Center Director); Memphis, TN
- Andrew Glass, MD—Kaiser Permanente Northwest, Center for Health Research (Senior Investigator); Portland, OR
- Mark Gorman—Individual Consultant; Silver Spring, MD
- Michael J. Hassett, MD, MPH—Dana-Farber Cancer Institute, Outpatient Clinic and Treatment Center (staff Physician, Breast Oncology); Boston, MA
- Karl Lorenz, MD, MSHS—UCLA, Department of Veterans Affairs (Associate Professor); Rand Health; Los Angeles, CA
- Joan McClure, MS—National Comprehensive Cancer Network, Clinical Information and Publications (Senior Vice President); Fort Washington, PA
- Bruce Minsky, MD—MD Anderson Hospital, Department of Radiation Oncology (Professor and Director of Clinical Research); Houston, TX
- Shirley Stagner, MSN, ONP, AOCNP—Lawrence Hospital Center, Cancer Survivorship Program (Nurse Practitioner); Bronxville, NY
- Janet H. Van Cleave, PhD, MSN, AOCNP—New York University College of Nursing (Assistant Professor); New York, NY
- Sandra L. Wong, MD MS—University of Michigan Health System, Division of Surgical Oncology (Physician); Ann Arbor, MI

In addition to the TEP, the team consulted with a work group consisting of representatives from each of the 11 PPS-exempt cancer hospitals. The purpose of this work group was to understand the quality measurement and improvement activities taking place at the PPS-exempt cancer hospitals and to learn their perspective in the importance and usefulness of measures we were evaluating for the program, including this measure.

PPS-exempt Cancer Hospital Measure Development Workgroup:

- J. Robert Beck, MD—American Oncologic Hospital (Fox Chase) (Senior Vice President and Chief Academic Officer)
- Charles Borden, MBA—Arthur G. James Cancer Hospital and Research Institute (Associate Executive Director of Quality and Patient Safety)
- Joe Jacobson, MD—Dana Farber Cancer Institute (Chief Quality officer)
- Barbara Jagels, MHA, RN, OCN—Seattle Cancer Care Alliance (Fred Hutchinson Cancer Research Center) (Director of Nursing and Clinical Excellence)
- Dana Jenkins—Roswell Park Memorial Institute (Vice President of Organizational Improvement)
- Tricia Kassab, RN, MS, CPHQ, HACP—City of Hope National Medical center (Vice President of Quality and Patient Safety)
- Jeremy Miransky, PhD—Memorial Hospital for cancer and Allied Disease (MSKCC) (Quality Analytics Manager)
- Shyroll Morris—University of Miami Hospital and Clinics
- Thomas Ross, MS—H. Lee Moffitt Cancer and research Institute Hospital, Inc. (Director of Quality and Safety)
- Anthony Senagore, MD—University of Southern California Kenneth Norris Jr. Cancer Hospital (Chief of Colorectal Surgery)
- Ron Walters, MD, MHA, MBA—The University of Texas MD Anderson Cancer Center (Associate Vice Present of Medical Operations and Informatics)
- Saul Weingart, MD, PhD—Dana Farber Cancer Institute (Vice Present for Quality Improvement and Patient Safety)

A subset of the TEP supplemented by representatives for the PPS-Exempt Cancer Hospitals supported the refinement and maintenance of the measure.

Expert Working Group Members:

- Peter B. Bach, MD, MAPP—Memorial Sloan-Kettering Cancer Center, Center for Health Policy and Outcomes (Director); New York, NY
- Stephen Edge, MD—Baptist Cancer Center (Cancer Center Director); Memphis, TN
- Michael J. Hassett, MD, MPH—Dana-Farber Cancer Institute, Outpatient Clinic and Treatment Center (staff Physician, Breast

Oncology); Boston, MA
- Karl Lorenz, MD, MSHS—UCLA, Department of Veterans Affairs (Associate Professor); Rand Health; Los Angeles, CA
- Allison Snyderman, PhD—Memorial Sloan-Kettering Cancer Center, Center for Health Policy and Outcomes (Researcher); New York, NY
- Tracy Spinks, CPH—MD Anderson Hospital, Cancer Care Delivery (Program Director); Houston, TX

The CORE/Mathematica measure development team met regularly during development and is comprised of experts in internal medicine, quality outcomes measurement, and measure development.

CORE/Mathematica Measure Development Team:
- Faseeha Altaf, MPH—Research Project Coordinator, CORE
- Liz Crane, BA—Research Assistant, Mathematica
- Thomas Croghan, MD—Clinical Investigator, Mathematica
- Hayley Dykhoff, BA – Research Assistant, CORE
- Lori Geary, MD, SM—Associate Director, CORE
- Angela Merrill, PhD—Senior Researcher, Mathematica
- Julia Montague, MPH – Project Manager, CORE
- Craig Parzynski, MS—Lead Analyst, CORE
- Chris Rankin—Programmer, Mathematica
- Christine Holland, MA, PMP—Researcher, Mathematica
- Joseph S. Ross, MD, MHS—Clinical Investigator, CORE
- Lori Schroeder, LLM, JD – Project Manager, CORE
- Fei Xing, PhD—Statistician, Mathematica

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:**

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?** Not applicable

Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable

Ad.7 **Disclaimers:** Not applicable

Ad.8 **Additional Information/Comments:** Not applicable

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 #: 2963
Measure Title: Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients
Measure Steward: PCPI
Brief Description of Measure: Percentage of patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy who did not have a bone scan performed at any time since diagnosis of prostate cancer
Developer Rationale: Multiple studies have indicated that a bone scan is not clinically necessary for staging prostate cancer in men with a low risk of recurrence and receiving primary therapy. For patients who are categorized as low-risk, bone scans are unlikely to identify their disease. Furthermore, bone scans are not necessary for low-risk patients who have no history or if the clinical examination suggests no bony involvement. Less than 1% of low-risk patients are at risk of metastatic disease.

While clinical practice guidelines do not recommend bone scans in low-risk prostate cancer patients, overuse is still common. An analysis of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it (1). The analysis also found that the use of bone scans in low-risk patients leads to an annual cost of $4 million dollars to Medicare. The overuse of bone scan imaging for low-risk prostate cancer patients is a concept included on the American Urological Association's list in the Choosing Wisely Initiative as a means to promote adherence to evidence-based imaging practices and to reduce health care dollars wasted (2). This measure is intended to promote adherence to evidence-based imaging practices, lessen the financial burden of unnecessary imaging, and ultimately to improve the quality of care for prostate cancer patients in the United States.

Citations:

1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

2. American Urological Association. A routine bone scan is unnecessary in men with low-risk prostate cancer. Choosing Wisely Initiative. Released February 21, 2013. Accessed February 25, 2016.

Numerator Statement: Patients who did not have a bone scan performed at any time since diagnosis of prostate cancer
Denominator Statement: All patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence, receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy
Denominator Exclusions: Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons)

Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician)

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

This measure is the new eMeasure version of measure 0389. The information provided for Evidence and Opportunity for Improvement is identical to that submitted for 0389. Measure 0389 will be discussed first – the ratings for evidence and opportunity for improvement will automatically be assigned to this eMeasure without further discussion.

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**

The evidence provided for this new e-Measure is the same as the evidence provided for the existing claims-based measure, #0389.

Summary of prior review in 2012 for #0389:

- The developer provided a best practice statement, a clinical practice guideline, and a systematic review of the body of evidence to demonstrate the use of bone scans for low risk prostate cancer patients is not supported by the evidence, is extremely costly, and unnecessarily exposes patients to radiation.
- The Prostate-Specific Antigen Best Practice Statement: 2009 Update from American Urological Association (AUA) recommended:
  - Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL**. Level of evidence: No grade assigned**
- The National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011 recommended:
  - For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors or symptomatic disease**. Level of evidence: NCCN grade 2A** (2A is defined as the recommendation is based on **lower-level evidence** and there is uniform NCCN consensus)
- The systematic review was associated with the development of the best practice statement and clinical guideline; the developer cites the number of studies associated with each though the details of the Quality, Quantity, and Consistency of the evidence was not provided.

Updates:

- The AUA Best Practice Statement and NCCN guideline were updated in 2013 and 2016, respectively. There were no changes to the recommendations since the previous submission.
- The developer provided new evidence: ACR Appropriateness Criteria. Prostate cancer- pretreatment detection, staging, and surveillance. American College of Radiology. 2012
  - ⋯only patients with a PSA ≥20 ng/ml (with any T stage or Gleason score), locally advanced disease (T3 or T4 with any PSA or Gleason score), or Gleason score ≥8 (with any PSA or T stage) should be considered for a radionuclide bone scan [91,99,101]. Patients with skeletal symptoms or advanced-stage disease should also be considered candidates for bone scans. p. 7. **Level of evidence based on the ACR 2012 Appropriateness Criteria:** 8 - Usually appropriate (Rating Scale: 1,2,3 Usually not appropriate;

4,5,6 May be appropriate; 7,8,9 Usually appropriate)

Exception to evidence

N/A

Guidance from the Evidence Algorithm: Process measure/systematic review (Box 3) → Specific information on QQC not presented (Box 4) → New evidence rated as 'usually appropriate'(Box 6) → Moderate (without QQC from SR, MODERATE is highest potential rating)

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

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

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

- The developer provided the following PQRS EHR, Registry, and Part B Claims data from January 1, 2014 – December 31, 2014:

	EHR Performance Rate	Registry Performance Rate
Mean	90.76%	90.24%
Minimum	50.0%	0.00%
Maximum	100.0%	100.00%

PQRS Experience Report

	Average Performance Rates
2010	71.60%
2011	90.50%
2012	92.50%
2013	88.50%

- NQF asks for performance scores on the measure as specified (current and over time). The developer provided registry/PQRS performance data. It is not clear if the EHR performance rates provided were obtained from the eMeasure specifications.
- NQF also asks that when providing performance scores, the following following information be provided: mean, standard deviation, min, max, interquartile range, scores by decile, and a description of the data source (number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).
- The developer provided additional data from the literature on prostate cancer patients who received a bone scan although it was not recommend.

Disparities:

- The developer did not provide any data on disparities from the measure as specified.
- The developer stated that while this measure is included in federal reporting programs, those programs have not yet made disparities data available to analyze and report.
- The developer provided a summary of data from the literature that compares the incidence, prevalence, and death rates between African American men and white men due to prostate cancer.

Questions for the Committee:

- *Without data from the eMeasure as specified, do you agree that the performance rates for the eMeasure are the same as the performance rates from the registry/PQRS?*
- *Does the data presented adequately demonstrate a quality problem and opportunity for improvement?*
- *Does the data presented demonstrate 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**

Rationale: Insufficient information provide to determine if a performance gap exists.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): electronic clinical data, electronic health record (EHR). This is an eMeasure.

Specifications:

- HQMF specifications for eMeasure are included in the document set on SharePoint. See eMeasure Technical Review below.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this eMeasure can be consistently implemented?*

eMeasure Technical Advisor(s) review (if not an eMeasure, delete this section):

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from both a simulated data set and an electronic healthcare records system demonstrating the measure logic can be interpreted precisely and unambiguously
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with the measure developer indicates that the measure logic is feasible. This is a legacy eMeasure included in the Meaningful Use program. The developer submitted Bonnie testing results to establish feasibility, and provided an interpretation of the testing process and results in the measure testing attachment.

2a2. Reliability Testing Testing attachment

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.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing:

- The dataset used for testing included 34 synthetic patients created in the Bonnie testing system simulating the year 2012. The developer stated that data from the Meaningful Use program are currently unavailable.
- Data element validity testing was performed and will count for data element reliability as well – see validity testing section.
- The developer provided reliability results from the registry/claims-based version of the eMeasure and stated that, "Once data are available for analysis of the eCQM, we expect that the reliability test results will be comparable."

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →Empirical reliability testing (Box 2) →Empirical validity testing of patient-level data (Box 3) →Refer to validity testing of patient-level data elements using Bonnie tool (Box 10) →Method appropriate for legacy eMeasures (Box 11) → Moderate (highest eligible rating is MODERATE)

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results from the Bonnie tool demonstrate sufficient reliability so that differences in performance can be identified?*
- *Do you agree that the reliability test results of the eMeasure will be comparable to the registry/claims-based measure?*

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

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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.

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- The Bonnie testing tool and environment with 34 synthetic patient records were used to test the measure logic and value sets. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure.
- The numerator, denominator, exceptions , and logic statement were tested to confirm actual results met

expectations.

- The developer also conducted a review of the measure specifications to confirm the measure logic was properly expressed with the current version of the QDM and that the logic matches the clinical intent of the measure.
- Face validity was assessed using a panel of 17 experts from representation from the PCPI Measures Advisory Committee.

Validity testing results:

- The testing results from the Bonnie tool reached 100% coverage and confirmed there was a test case for each pathway of logic (negative and positive test cases).
- The measure also had a 100% passing rate which confirmed that all the test cases performed as expected.
- **80%** (10) of the respondents from the PCPI Measures Advisory Committee either agreed or strongly agreed that this measure can accurately distinguish good and poor quality.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results from the Bonnie tool demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exceptions in the eCQM are harmonized with the registry version of this measure (#0389) and include:
  - Documentation of reason(s) for performing a bone scan (including documented pain, salvage therapy, other medical reasons, bone scan ordered by someone other than reporting physician)
- The developer provided the results from the exceptions analysis conducted on the registry/claims-based measure (#0389).

Questions for the Committee:

- *Is it reasonable to assume that the impact of the exclusions will be similar for the eMeasure and the registry/claims-based measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provided measures of central tendacy, variability, and dispersion from the registry/claims-based measure.

Question for the Committee:

- *Does the Committee agree the eMeasure will demonstrate similar results to the registry/claims-based measure?*

2b6. Comparability of data sources/methods:

- N/A

2b7. Missing Data

- The developer stated that data are not available to complete testing.

Guidance from validity algorithm: Specifications consistent with evidence (Box 1)→Some threats to validity addressed (Box 2) →Empirical validity testing (Box 3) → Face validity testing (Box4) and empirical testing of data elements using Bonnie tool (Box 10) →Method appropriate for legacy eMeasures (Box 11) →Moderate (highest eligible rating is MODERATE)

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

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

- The developer provided information on feasibility testing in the eMeasure Feasbility Score Card for two implementation sites and an explanation for scores below 2. However, the developer did not identify the two EHRs used for feasibility testing.
- Bonnie testing verified that the measure logic is functional, but not all of the required data elements exists as structured data in the unidentified EHRs that were used for testing feasibility. However, the developer provided EHR performance rates in section 1b, therefore it is assumed the eMeasure is feasible for some users.

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?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

4. Usability and 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details:

- This measure is in Meaningful Use Stage 2 (EHR Incentive Program) sponsored by CMS. The Medicare and Medicaid EHR Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology. Eligibility for incentive payments for the "meaningful use" of certified EHR technology is established if all program requirements are met, including successful implementation and reporting of program measures, which include this measure, to demonstrate meaningful use of EHR technology.
  - The eMeasure version of this measure is not publicly reported or used in PQRS.

Improvement results:

- The developer included the performance rates previously reported in 1b.2. Progress on improvement, including trends in performance results, number and percentage of people receiving high-quality healthcare, geographic area and number and percentage of accountable entities and patients were not discussed.

Potential harms:

- The developer stated that they are not aware of any unintended consequences at this time, but take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

Feedback: In 2012, public and member comments received on the registry/claims-based measure included:

- Commenters indicated that the Steering Committee should consider clarifying 'low risk' status for the measure population and that classification for measurement purposes should be based on staging information available at the time of decision making regarding whether or not to order a bone scan.
- Commenters believed that the measure should clearly articulate that even those patients with a positive bone scan remain in the denominator of this measure, even though the bone scan ultimately demonstrates that they are not actually low risk.
- Comments reflected questions on the measure specifications, specifically:
    - It is unclear how treatment interplays with this measure.
    - The numerator captures patients who did not have a 'bone scan performed prior to initiation of treatment nor at any time since diagnosis.
    - Patient eligibility for the denominator should be based on criteria known before the decision to deliver the service (the bone scan) is considered.
    - Exclusion criteria (i.e. treatment planned for future, patient preference, vulnerable health status, and poor access to care)
- Several commenters supported this measure.
- The measure developer's response was:
- The AUA/AMA-PCPI Prostate Cancer Work Group appreciates your comment. The Work Group will consider your feedback about the risk stratification, when the measure undergoes formal review and maintenance, according to the AMA-PCPI measure development/maintenance methodology, in the future. Additionally, the measure contains a medical exception, which allows physicians to use clinical judgment in order to have a bone scan performed on those low-risk prostate cancer patients who have a medical reason documented.
- The denominator was constructed so any patient that has already been stratified as a low risk patient and is being treated according to the low risk strata would be captured in the measure. The measure is aiming to reduce the use of bone scans that are clinically unnecessary, in low risk patients who generally have no indication for imaging studies. Additionally, the measure contains a medical exception, which allows physicians to use clinical judgment in order to n performed on those low-risk prostate cancer patients who have a medical reason documented.

***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

Criterion 5: Related and Competing Measures

Related or competing measures:

- 0390 : Prostate Cancer: Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients
- 1853 : Radical Prostatectomy Pathology Reporting

Harmonization:

- 0390 and 1853 measure different target populations and addresses different aspects of prostate cancer care

Pre-meeting public and member comments

- The American Urological Association supports the continued use of this important measure.

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0389 **NQF Project:** Cancer Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The process of identifying the patient´s risk strata prior to ordering any imaging studies is related to improved outcomes, including cost reduction and reduction of radiation exposure.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The evidence directly supports the specified measure. The measure specifically identifies the risk strata for whom bone scans are inappropriate. The guideline and best practice statement do not recommend bone scans for patients included in the low or very low risk strata.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: The AUA best practice statement references a systematic review of 23 studies, examining the utility of bone scan.

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

The description of the evidence review in the NCCN guideline did not address the overall quantity of studies in the body of evidence. However, 223 articles are cited in NCCN´s prostate cancer guideline´s reference section.

AUA 2013 Guideline:
The guideline cites that 23 studies were reviewed to develop the recommendation statement.

NCCN 2016 Guideline:
Information regarding the total number of studies and type of study designs included in the body of evidence is not available. However, the guideline cites 1 observational study in support of the recommendation statement.

ACR 2012 Appropriateness Criteria:
The guideline cites 6 observational diagnostic studies and 3 reviews/other diagnostic studies in support of the recommendation statement.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The systematic review of the literature (cited within the AUA best practice statement) states the following:

Studies were eligible only if newly diagnosed PC cases with no previous management were included. Studies were excluded if details regarding the patient population or results were significantly lacking.

The findings of this study are based on data pooled primarily from retrospective series. It is possible that inherent biases occurred in reporting. For example although not always reported, it may have been that in some studies those cases with a positive CT did not proceed to surgery. Results from these cases may not have been reported, thereby lowering the apparent detection rate in the reported population.

As with all studies of this nature, there are limitations to the findings. Unfortunately, not all series reported results based on the pooled groupings of PSA, tumor stage and Gleason score used herein. In addition, not all studies graded disease using Gleason score. As a result, inclusion of data from all cases was not justified. Fortunately, large patient numbers were remaining to allow for small confidence intervals around estimates.

Because of the nature of this study, it is not possible to make recommendations based on combinations of prognostic factors. For example bone scanning detected metastases in 6.4% of patients with localized disease but it is not possible to tell what proportionwere at risk by virtue of increased PSA or Gleason score, for which scanning would have been indicated. Presumably, some of those patients with positive bone scans would have been at risk from either of these factors. Therefore, it could be argued that the true risk for patients with low PSA, low Gleason score and localized disease is less than those numbers reported here. Also, most of these studies were published in the 1990s and contained results for patients seen before the widespread use of PSA screening. Therefore, no distinction can be made in patients with organ confined disease between those with palpable and nonpalpable tumors. Again, it could be argued that due to stage migration within this group, numbers reported here are higher than the true risk.

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

The quality of the body of evidence supporting the NCCN guideline recommendation is summarized according to the NCCN categories of evidence and consensus as being based on "lower-level evidence."

AUA 2013 Guideline:
The quality of body of evidence was not included.

NCCN 2016 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence, however, the guideline is a Category 2A: "Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate."

ACR Appropriateness Criteria:
The quality of body of evidence was not included.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The systematic review referenced by the AUA best practice statement does not contain information about consistency of results across

studies.

Although there is no explicit statement regarding the overall consistency of results across studies in the NCCN guideline, the recommendation received uniform NCCN consensus that the recommendation is appropriate.

AUA 2013 Guideline:
The consistency of results across studies was not reviewed.

NCCN 2016 Guideline:
The guideline does not provide the consistency of results across studies, however, the recommendation received uniform NCCN consensus that the intervention is appropriate.

ACR Appropriateness Criteria:
The consistency of results across studies was not reviewed.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**
Overuse of bone scans among prostate cancer patients is extremely costly and unnecessarily exposes patients to radiation. The use of bone scans for low risk prostate cancer patients is not supported by evidence.

AUA 2013 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence. However, they do include the following summary information regarding the benefits of not performing a bone scan on low-risk prostate cancer patients:
"The authors concluded that low-risk patients are unlikely to have disease identified by bone scan. Accordingly, bone scans are generally not necessary in patients with newly diagnosed prostate cancer who have a PSA < 20 ng/mL unless the history or clinical examination suggests bony involvement."

While no harms were mentioned, it is expected that the harms of using a bone scan on low-risk patients includes unnecessary exposure to radiation which can have potential harms such as radiation burns, adverse reactions to contrast media, and radiation-induced malignancy.

NCCN 2016 Guideline:
The guideline does not include an overall estimate of benefit from the body of evidence. However, the guideline does state "Patients with low-and intermediate-risk and low postoperative serum PSA levels have a very low risk of positive bone scans or CT scans. In a series of 414 bone scans performed in 230 men with biochemical recurrence after radical prostatectomy, the rate of a positive scan for men with PSA >10 ng/mL was only 4%."

While the guidelines did not describe how the harms studied affected net benefits, they did state that the risks of imaging include adverse reaction to contrast media, false-positive scans, and over-detection. Additional risks of imaging include radiation burns, cataracts, radiation-induced malignancy, and adverse reactions to contrast material delivered by intravenous, oral, or rectal routes.

ACR 2012 Appropriateness Criteria:
The guideline does not include an overall estimate of benefit from the body of evidence. However, they do include the following summary information regarding the benefits of not performing a bone scan on low-risk prostate cancer patients:
"Work by Oesterling and others has shown that in patients with low PSA level (<10 ng/ml) who have no pain, the yield of a staging bone scan is too low to warrant its routine use. In their experience, no patient with a PSA ≤ 10 ng/ml had a positive bone scan and only one patient in 300 with a PSA level ≤ 20 ng/ml had a positive radionuclide scan"

While no harms were mentioned, it is expected that the harms of using a bone scan on low-risk patients includes unnecessary exposure to radiation which can have potential harms such as radiation burns, adverse reactions to contrast media, and radiation-induced malignancy.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

AUA 2013 Guideline:

No grade has been assigned for the quality of evidence.

NCCN 2016 Guideline:
Yes

ACR 2012 Appropriateness Criteria:
Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: NCCN Prostate Cancer Panel

Andrew J. Armstrong, MD, ScM
Robert R. Bahnson, MD
Barry Boston, MD
J. Erik Busby, MD
Anthony Victor D´Amico, MD, PhD
James A. Eastham, MD
Charles A. Enke, MD
Thomas A. Farrington
Lauren Gallagher, RPh, PhD
Kristina M. Gregory, RN, MSN, OCN
Celestia S. Higano, MD, FACP
Maria Ho, PhD
Eric Mark Horwitz, MD
Philip W. Kantoff, MD
Mark H. Kawachi, MD
Michael Kuettel, MD, MBA, PhD
Richard J. Lee, MD, PhD
Gary R. MacVicar, MD
Arnold W. Malcolm, MD, FACR
Joan S. McClure, MS
David Miller, MD, MPH
James L. Mohler, MD
Elizabeth R. Plimack, MD, MS
Julio M. Pow-Sang, MD
Mack Roach, MD
Eric Rohren, MD, PhD
Stan Rosenfeld
Dorothy Shead, MS
Sandy Srinivas, MD
Seth A. Strope, MD, MPH
Jonathan Tward, MD, PhD
Przemyslaw Twardowski, MD
Patrick C. Walsh, MD

The NCCN Guidelines are updated at least annually in an evidence-based process integrated with the expert judgment of multidisciplinary panels of expert physicians from NCCN Member Institutions. NCCN depends on the NCCN Guidelines Panel Members to reach decisions objectively, without being influenced or appearing to be influenced by conflicting interests.

All panel member disclosures are available at www.nccn.org.

NCCN 2016 Guideline Prostate Cancer Panel:
James Mohler, MD
Andrew Armstrong, MD

Robert Bahnson, MD
Anthony Victor D'Amico, MD PhD
Brian Davis, MD, PhD
James Eastham, MD
Charles Enke, MD
Thomas Farrington,
Celestia Higano, MD
Eric Horwitz, MD
Michael Hurwitz, MD, PhD
Christoper Kane, MD
Mark Kawachi, MD
Michael Kuettel, MD, MBA, PhD
Richard Lee, MD, PhD
Joshua Meeks, MD, PhD
Daivd Penson, MD, MPH
Elizabeth Plimack, MD, MS
Julio Pow-Sang, MD
David raben, MD
Sylvia Richey, MD
March Roach, III, MD
Stan Rosenfeld
Edward Schaeffer, MD, PhD
Ted Skolarus, MD
Eric Small, MD
Guru Sonpavde, MD
Sandy Srinivas, MD
Seth Strope, MD, MPH
Johnathon Tward, MD, PhD

All panel member disclosures are available at www.nccn.org.

ACR 2012 Appropriateness Criteria: Expert Panels on Urologic Imaging and Radiation Oncology–Prostate:
Steven C. Eberhardt, MD
Scott Carter, MD
David D. Casalino, MD
Gregory Merrick, MD
Steven J. Frank, MD
Alexander R. Gottschalk, MD, PhD
John R. Leyendecker, MD
Paul L. Nguyen, MD
Aytekin Oto, MD
Christopher Porter, MD
Erick M. Remer, MD
Seth A. Rosenthal, MD

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major

disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guideline:
NCCN Categories of Evidence and Consensus
Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

ACR 2012 Appropriateness Criteria:
Study Quality Category Definitions
- Category 1 The study is well-designed and accounts for common biases.
- Category 2 The study is moderately well-designed and accounts for most common biases.
- Category 3 There are important study design limitations.
- Category 4 The study is not useful as primary evidence. The article may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus. For example:

a) the study does not meet the criteria for or is not a hypothesis-based clinical study (e.g., a book chapter or case report or case series description);
b) the study my synthesize and draw conclusions about several studies such as a literature review article or book chapter but is not primary evidence;
c) the study is an expert opinion or consensus document.

1c.13 Grade Assigned to the Body of Evidence: No grade for AUA best practice statement, NCCN grade 2A

AUA 2013 Guideline:
No grade has been assigned to the body of evidence.

NCCN 2016 Guideline:
Level of evidence assigned: Category 2A

ACR 2012 Appropriateness Criteria:
The guideline includes 5 observational diagnostic studies with a study quality of 3 and 1 observational diagnostic study with a study quality of 2. The guideline also includes 3 reviews/other diagnostic studies with a study quality of 4.

1c.14 Summary of Controversy/Contradictory Evidence: No contradictory evidence has been identified.

AUA 2013 Guideline:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

NCCN 2016 Guideline:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

ACR 2012 Appropriateness Criteria:
The review does not provide a summary of controversy/contradictory evidence. However, the potential harms expected are referred to in section 1c8.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
A radionuclide bone scan is traditionally the first examination obtained. If the bone scan is positive for metastatic disease, no further

imaging studies are
necessary. If it is inconclusive, further imaging studies are performed, including conventional radiographs, MRI, or computed tomography (CT). However, the level of posttreatment PSA that should prompt a bone scan is uncertain. In a study of patients with biochemical failure following radical prostatectomy, the probability of a positive bone scan was less than 5% with PSA levels between 40-45 ng/ml. In another study, bone scan was limited until PSA rose above 30-40 ng/ml. Men with a PSADT of <6 months after radical prostatectomy were at increased risk of a positive bone scan (26% vs 3%) or
positive CT (24% vs 0%) compared to those with longer PSADT. Kane et al reported that most patients with a positive bone scan had a high PSA level (mean of 61.3 ng/ml) and a high PSA velocity (>0.5 ng/ml/month).

American College of Radiology. ACR Appropriateness Criteria. Post-treatment Follow-up of Prostate Cancer. 2011. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/app_criteria/pdf/ExpertPanelonUrologicImaging/PostTreatmentFollowUpofProstateCancerDoc10.aspx

The results of a retrospective review demonstrate extensive overuse of bone scan imaging among VA patients with low-risk prostate cancer.Overall, the rate of bone scan imaging among men with low-risk features was 25% with no positive findings.

Palvolgyl R, Daskivich TJ, Chamie K, Kwan L, Litwin MS. Bone Scan Overuse in Staging of Prostate Cancer: An Analysis of a Veterans Affairs Cohort.

Citation for the systematic review of literature, referenced in sections 1c.5 and 1c.6 is below:

Abuzallouf S, Dayes I, Lukka H. Baseline Stagin of Newly Diagnosed Prostate Cancer: A Summary of the Literature. Journal of Urology 2004;171:2122-2127.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
1. Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL.

2. For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors or symptomatic disease.

AUA 2013 Guideline:
Routine use of a bone scan is not required for staging asymptomatic men with clinically localized prostate cancer when their PSA level is equal to or less than 20.0 ng/mL. p. 4

NCCN 2016 Guideline:
For symptomatic patients and/or those with a life expectancy of greater than 5 years, a bone scan is appropriate for patients with any of the following: 1) T1 disease with PSA over 20 ng/mL or T2 disease with PSA over 10 ng/mL; 2) a Gleason score of 8 or higher; 3) T3 to T4 tumors; or 4) symptomatic disease. (Category 2A) p. 64

ACR 2012 Appropriateness Criteria:

Clinical Condition: Prostate Cancer — Pretreatment Detection, Staging, and Surveillance

Variant 3: Prostate cancer diagnosed on biopsy, patient at high risk for locally advanced disease and metastases (AJCC Groups III and IV). Example: PSA ≥20 or Gleason 8-10 or clinical stage T2c or higher.

Radiologic Procedure	Rating	Comments	RRL*
MRI pelvis without and with contrast	8	Should include dynamic contrast-enhanced (DCE) technique. See statement regarding contrast in text under "Anticipated Exceptions."	O
Tc-99m bone scan whole body	8		☢☢☢
CT abdomen and pelvis with contrast	7		☢☢☢☢
MRI pelvis without contrast	6		O
CT abdomen and pelvis without contrast	6	If contrast contraindicated.	☢☢☢☢
X-ray area of interest	4	Appropriate if bone scan or symptoms suggest possible involvement.	Varies
FDG-PET/CT whole body	4		☢☢☢☢
CT abdomen and pelvis without and with contrast	2		☢☢☢☢
In-111 capromab pendetide scan	2		☢☢☢☢
Rating Scale: 1,2,3 Usually not appropriate; 4,5,6 May be appropriate; 7,8,9 Usually appropriate			*Relative Radiation Level

…only patients with a PSA ≥20 ng/ml (with any T stage or Gleason score), locally advanced disease (T3 or T4 with any PSA or Gleason score), or Gleason score ≥8 (with any PSA or T stage) should be considered for a radionuclide bone scan [91,99,101]. Patients with skeletal symptoms or advanced-stage disease should also be considered candidates for bone scans. p. 7

1c.17 Clinical Practice Guideline Citation: 1. Prostate-Specific Antigen Best Practice Statement: 2009 Update from American Urological Association, American Urological Association Education and Research, Inc. Available at:http://www.auanet.org/content/guidelines-and-qualitycare/clinical-guidelines/main-reports/psa09.pdf.

2. National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Prostate Cancer. Version 4.2011. Available at www.nccn.org

AUA 2013 Guideline:
Carroll P, Albertsen PC, Greene K, et al. American Urological Association Education and Research, Inc. PSA testing for the pretreatment staging and posttreatment management of prostate cancer: 2013 Revision of 2009 Best Practice Statement. Linthicum, MD: American Urological Association Education and Research, Inc. 2013. Available at: https://www.auanet.org/common/pdf/education/clinical-guidance/Prostate-Specific-Antigen.pdf

NCCN 2016 Guideline:
National Comprehensive Cancer Network (NCCN). Clinical practice guidelines in oncology: prostate cancer. Version 2.2016. Available at www.nccn.org

ACR 2012 Appropriateness Criteria
Eberhardt SC, Carter S, Casalino D, et al. ACR Appropriateness Criteria. Prostate cancer- pretreatment detection, staging, and surveillance. American College of Radiology. 2012. Available at: https://acsearch.acr.org/list

1c.18 National Guideline Clearinghouse or other URL: http://www.auanet.org/content/media/psa09.pdf and www.nccn.org

AUA 2013 Guideline:
Available at http://www.auanet.org/education/aua-guidelines.cfm

NCCN 2016 Guideline:

Available at NCCN.org

ACR 2012 Appropriateness Criteria
Available at: https://acsearch.acr.org/list

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

AUA 2013 Guideline:
No

NCCN 2016 Guideline:
Yes

ACR 2012 Appropriateness Criteria
Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: NCCN Prostate Cancer Panel is listed in section 1c.10

NCCN 2016 Guideline:
NCCN Prostate Cancer Panel is listed in section 1c.10

ACR 2012 Appropriateness Criteria
Expert Panels on Urologic Imaging and Radiation Oncology–Prostate is listed in section 1c.10

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: NCCN Categories of Evidence and Consensus

Category 1: The recommendation is based on high-level evidence (e.g. randomized controlled trials) and there is uniform NCCN consensus.

Category 2A: The recommendation is based on lower-level evidence and there is uniform NCCN consensus.

Category 2B: The recommendation is based on lower-level evidence and there is nonuniform NCCN consensus (but no major disagreement).

Category 3: The recommendation is based on any level of evidence but reflects major disagreement.

NCCN 2016 Guideline:
NCCN Categories of Evidence and Consensus
Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate

Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.

Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.

ACR 2012 Appropriateness Criteria:
ACR Appropriateness Criteria Methodology
The ACR AC methodology is based on the RAND Appropriateness Method2. The appropriateness ratings for each of the procedures or treatments included in the AC topics are determined using a modified Delphi method. A series of surveys are conducted to elicit each panelist's expert interpretation of the evidence, based on the available data, regarding the appropriateness of an imaging or therapeutic procedure for a specific clinical scenario. The expert panel members review the evidence presented and assess the risks or harms of

doing the procedure balanced with the benefits of performing the procedure. The direct or indirect costs of a procedure are not considered as a risk or harm when determining appropriateness. When the evidence for a specific topic and variant is uncertain or incomplete, expert opinion may supplement the available evidence or may be the sole source for assessing the appropriateness.

The appropriateness is represented on an ordinal scale that uses integers from 1 to 9 grouped into three categories: 1, 2, or 3 are in the category "usually not appropriate" where the harms of doing the procedure outweigh the benefits; and 7, 8, or 9 are in the category "usually appropriate" where the benefits of doing a procedure outweigh the harms or risks. The middle category, designated "may be appropriate", is represented by 4, 5, or 6 on the scale. The middle category is when the risks and benefits are equivocal or unclear, the dispersion of the individual ratings from the group median rating is too large (i.e., disagreement), the evidence is contradictory or unclear, or there are special circumstances or subpopulations which could influence the risks or benefits that are embedded in the variant.

The ratings assigned by each panel member are presented in a table displaying the frequency distribution of the ratings without identifying which members provided any particular rating. To determine the panel's recommendation, the rating category that contains the median group rating without disagreement is selected. This may be determined after either the first or second rating round. If there is disagreement after the second rating round, the recommendation is "May be appropriate."

This modified Delphi method enables each panelist to articulate his or her individual interpretations of the evidence or expert opinion without excessive influence from fellow panelists in a simple, standardized, and economical process.

1c.23 Grade Assigned to the Recommendation: No grade for AUA best practice statement, NCCN grade 2A

AUA 2013 Guideline:
No grade or definition has been provided for this guideline.

NCCN 2016 Guideline:
Level of evidence assigned: Category 2A

ACR 2012 Appropriateness Criteria:
Appropriateness Rating Assigned: 8

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: Moderate **1c.26** Quality: Moderate **1c.27** Consistency: Moderate

1. Evidence, Performance Gap, Priority – 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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

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

0389_Evidence_MSF5.0_Data-635278494960508026-635932939997439723-635948416107391693.doc

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., the benefits or improvements in quality envisioned by use of this measure)*

Multiple studies have indicated that a bone scan is not clinically necessary for staging prostate cancer in men with a low risk of recurrence and receiving primary therapy. For patients who are categorized as low-risk, bone scans are unlikely to identify their disease. Furthermore, bone scans are not necessary for low-risk patients who have no history or if the clinical examination suggests no bony involvement. Less than 1% of low-risk patients are at risk of metastatic disease.

While clinical practice guidelines do not recommend bone scans in low-risk prostate cancer patients, overuse is still common. An analysis of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it (1). The analysis also found that the use of bone scans in low-risk patients leads to an annual cost of $4 million dollars to Medicare. The overuse of bone scan imaging for low-risk prostate cancer patients is a concept included on the American Urological Association's list in the Choosing Wisely Initiative as a means to promote adherence to evidence-based imaging practices and to reduce health care dollars wasted (2). This measure is intended to promote adherence to evidence-based imaging practices, lessen the financial burden of unnecessary imaging, and ultimately to improve the quality of care for prostate cancer patients in the United States.

Citations:

1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

2. American Urological Association. A routine bone scan is unnecessary in men with low-risk prostate cancer. Choosing Wisely Initiative. Released February 21, 2013. Accessed February 25, 2016.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. 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 included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI
Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 EHR Performance Rate:
Mean: 90.76%
Maximum: 100.00%
Minimum: 50.00%

01/01/2014 – 12/31/2014 Registry Performance Rate:
Mean: 90.24%
Minimum: 0.00%
Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients the last several years are as follows:
2010: 71.60%
2011: 90.50%
2012: 92.50%
2013: 88.50%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 8.2% of eligible professionals reported on Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients for claims, registry, and electronic health records. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

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.
An analysis conducted by Falchook and colleagues of prostate cancer patients in the SEER-Medicare database diagnosed from 2004-2007 found that 43% of patients for whom a bone scan was not recommended received it. Given that low-risk patients have a less than 1% chance of developing a metastatic disease, the authors suggest that bone scan imaging in low-risk prostate cancer patients contributes to poor quality care and is a large contributor of health care dollars wasted in the United States (1). The literature recommends clinician education on guideline recommendations to spur improvement. These findings support the need for an NQF endorsed performance measure along with targeted initiatives to improve adherence to appropriate imaging.

Citation:
1. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

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 endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
While this measure is included in federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
African American/Black men have the highest incidence rate of prostate cancer in the United States and are more than twice as likely as white men to die from the disease (1). Between 2008-2012, the average annual prostate cancer incidence rate among African American men was 208.7 cases per 100,000 men, which was 70% higher than the rate in white men. Although prostate cancer incidence and mortality rates have been declining in African American and white men since 1991, the incidence, prevalence, and death rates remain comparably higher among African American men as compared to white men (2).

An analysis of the SEER Medicare database conducted by Falchook and colleagues found that imaging overuse was associated with nonwhite race, higher comorbidity, and regional education and income measures (3). An additional analysis SEER Medicare database found there was regional variation in the use of bone scans in low- and intermediate- risk patients with the highest use in the Northeast and lowest use in the West (4).

Citations:
1. National Cancer Institute. Cancer health disparities. http://www.cancer.gov/about-nci/organization/crchd/cancer-health-disparities-fact-sheet. Accessed February 12, 2016.

2. American Cancer Society Cancer Facts and Figures for African Americans 2016-2018. http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-047403.pdf. Accessed February 26, 2016.

3. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

4. Falchook AD, Salloum RG, Hendrix LH, Chen RC. Use of bone scan during initial prostate cancer workup, downstream procedures, and associated Medicare costs. Int J Radiat Oncol Biol Phys. 2014;89(2):243-248.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, High resource use

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

220,800 new cases of prostate cancer were diagnosed in 2015 (1). The number of new cases of prostate cancer was 137.9 per 100,000 men per year, based on age-adjusted cases and deaths between 2008-2012. In 2012, there were an estimated 2,795,592 men living with prostate cancer in the United State. Prostate cancer is the third most common type of cancer in the United State and accounts for 13.3% of all new cancer cases. Approximately 14% of men will be diagnosed with prostate cancer at some point in their lifetime, based on 2010-2012 data (2).

The annual estimated cost to Medicare of all bone scans for prostate cancer patients is $19,300,000 including $9,300,000 for low and intermediate-risk patients. An additional $2,000,000 is spent annually on downstream imaging bone scans for low- and- intermediate risk patients (3). The annual Medicare of bone scans for low-risk prostate cancer patients alone is $4,000,000 (4).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Siegel RL, Miller KD, Jemal A. Cancer statistics 2015. Ca Cancer J Clin. 2015;65:5-29.

2. SEER Cancer Statistics Factsheets: Prostate Cancer. National Cancer Institute. Bethesda, MD,http://seer.cancer.gov/statfacts/html/prost.html

3. Falchook AD, Salloum RG, Hendrix LH, Chen RC. Use of bone scan during initial prostate cancer workup, downstream procedures, and associated Medicare costs. Int J Radiat Oncol Biol Phys. 2014;89(2):243-248.

4. Falchook AD, Hendrix LH, Chen RC. Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer. J Oncol Pract. 2015. doi: 10.1200/JOP.2014.001818.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable. Not a PRO-PM.

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 subcriteria 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):*
Cancer, Cancer : Prostate

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse

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.)*
Measure specifications are included as an attachment with this submission. Additional measure details may be found at:http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/eCQM_Library.html Value sets at https://vsac.nlm.nih.gov

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 an eMeasure **Attachment:** EP_CMS129v6_NQF0389_PrCA_OveruseBoneScan_02182016-635948416097719507.zip

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:** EP_eCQM_ValueSets_CMS129v6_NQF0389_02182016-635948416089607351.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The measure description, denominator statement, denominator details and value sets, were revised based on updated risk strata to be consistent with NCCN guidelines.

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)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who did not have a bone scan performed at any time since diagnosis of prostate cancer

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Once for each procedure for treatment of prostate cancer (ie, interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy)during the 12-month reporting period

S.6. **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, 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.
For Registry:
To submit the numerator option for patients who did not have a bone scan performed at any time since diagnosis of prostate cancer, report the following CPT Category II code:

3270F – Bone scan not performed prior to initiation of treatment nor at any time since diagnosis of prostate cancer

For EHR Specifications:
HQMF eMeasure developed and is included in this submission.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All patients, regardless of age, with a diagnosis of prostate cancer at low (or very low) risk of recurrence, receiving interstitial prostate brachytherapy, OR external beam radiotherapy to the prostate, OR radical prostatectomy, OR cryotherapy

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, 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)*
Definitions:
Risk Strata Definitions: Very Low, Low, Intermediate, High, or Very High-
Very Low Risk - PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1c; AND presence of disease in fewer than 3 biopsy cores; AND <= 50% prostate cancer involvement in any core; AND PSA density <= 0.15 ng/mL/cm3.
Low Risk - PSA < 10 ng/mL; AND Gleason score 6 or less; AND clinical stage T1 to T2a.
Intermediate Risk - PSA 10 to 20 ng/mL; OR Gleason score 7; OR clinical stage T2b to T2c. Note: Patients with multiple adverse factors may be shifted into the high risk category.
High Risk - PSA > 20 ng/mL; OR Gleason score 8 to 10; OR clinically localized stage T3a. Note: Patients with multiple adverse factors may be shifted into the very high risk category.
Very High Risk - Clinical stage T3b to T4; OR primary Gleason pattern 5; OR more than 4 cores with Gleason score 8 to 10. (NCCN, 2016)

External beam radiotherapy - external beam radiotherapy refers to 3D conformal radiation therapy (3D-CRT), intensity modulated radiation therapy (IMRT), stereotactic body radiotherapy (SBRT), and proton beam therapy.

Note: Only patients with prostate cancer with low risk of recurrence will be counted in the denominator of this measure

For Registry:
Any male patient, regardless of age
AND
Diagnosis for prostate cancer (ICD-9-CM): 185
Diagnosis for prostate cancer (ICD-10-CM): C61
AND
Patient encounter during the reporting period (CPT): 55810, 55812, 55815, 55840, 55842, 55845, 55866, 55873, 55875, 77427, 77435, 77772, 77778, 77799
AND
Report the following CPT Category II Code to identify the risk of recurrence:
3271F: Low risk of recurrence, prostate cancer

For EHR:
HQMF eMeasure developed and is included in this submission.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Documentation of medical reason(s) for having a bone scan performed (including documented pain, salvage therapy, other medical reasons)

Documentation of system reason(s) for having a bone scan performed (including bone scan ordered by someone other than reporting physician)

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, 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)
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients, exceptions may include medical reason(s) (eg, documented pain, salvage therapy, other medical reasons) or system reason(s) (eg, bone scan ordered by someone other than reporting physician). Where examples of exceptions are included in the measure language, value sets for these examples are developed and included in the eMeasure. Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Additional details by data source are as follows:

For Registry:
Append modifier to CPT Category II code:
3269F with 1P - Documentation of medical reason(s) for performing a bone scan (including documented pain, salvage therapy, other medical reasons)

Append modifier to CPT Category II code:
3269F with 3P - Documentation of system reason(s) for performing a bone scan (including bone scan ordered by someone other than reporting physician)

For EHR:
HQMF eMeasure developed and is included in this submission.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer and have included these variables as recommended data elements to be collected.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment or risk stratification

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)

S.16. Type of score:
Rate/proportion

If other:

S.17. **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.18. **Calculation Algorithm/Measure Logic** *(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; aggregating data; risk adjustment; etc.)*
To calculate performance rates:
1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).
2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.
3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: medical reason(s) (eg, documented pain, salvage therapy, other medical reasons) or system reason(s) (eg, bone scan ordered by someone other than reporting physician)]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable. The measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable. The measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Patient eligibility is determined by a set of defined criteria relevant to a particular measure. If data required to determine patient eligibility are missing, those patients/cases would be ineligible for inclusion in the denominator and therefore the patient/case would be deleted.

If data required to determine if a denominator eligible patient qualifies for the numerator (or has a valid exclusion/exception) are missing, this case would represent a quality failure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*

IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Not applicable.

S.25. **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.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Clinician : Team

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinician Office/Clinic, Other
If other: Radiation Oncology Clinic/Department

S.28. **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. The measure is not a composite.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_2963_Avoidance_of_Overuse_of_bone_scans_Updated_4_1_2016.docx

Measure Number (*if previously endorsed*)**:** 2963

Measure Title: Prostate Cancer: Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients

Date of Submission: 3/11/2016

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

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.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: Bonnie Patient Test Deck	☒ other: Bonnie Patient Test Deck

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*).

Bonnie Patient Test Deck
This is an NQF measure that was previously endorsed for use in other modalities, but was retooled into an electronic clinical quality measure (eCQM). Per NQF guidance, regarding testing requirements for legacy measures, the Bonnie testing tool was used to evaluate the measure specifications, measure logic, and HQMF output, using synthetic patient information.

1.3. What are the dates of the data used in testing? The Bonnie test environment simulates the year 2012.

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: 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 has been retooled into an eCQM, it is included in both the PQRS program and the CMS Meaningful Use Stage 2 EHR Incentive Program. Data from the PQRS program included an insufficient sample size and were unable to be analyzed for reliability and data from the Meaningful Use program are currently unavailable. When EHR data becomes available, Signal to Noise Ratio analysis will be performed, to assess reliability.

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*)

34 synthetic patients were created in the Bonnie testing tool, to evaluate this measure. Patient information included in the testing tool include the following:

- Name

- Date of Birth
- Patient medical history
    - Diagnostic Studies
    - Diagnoses
    - Laboratory Tests
    - Procedures

Synthetic patients were created in order to ensure that the test deck included patients that pass,for each required data element. The 34 patients created to test this measure passed, as expected, and the measure logic performed, as expected, within the Bonnie tool.

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.

The Bonnie patient test deck was used for testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

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*)

Currently, there is insufficient performance data available for testing of the electronic clinical quality measure. However, claims and registry versions of this measure have been in use, as a part of national quality reporting programs since as early as 2008. T

The most recent PCPI testing project that was performed on measure #0389, the measure which was retooled into this eCQM, in order to evaluate the reliability of this measure was conducted via Signal to Noise Ratio analysis, using 2014 PQRS registry data.

Reliability of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

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*)

Signal to Noise Ratio analysis results

This measure has 0.84 reliability when evaluated at the minimum level of quality reporting events and 0.96 reliability at the average number of quality events.

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?*)

Signal to Noise Ratio analysis

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is very high.

Once data are available for analysis of the eCQM, we expect that the reliability test results will be comparable.

2b2. VALIDITY TESTING

2b2.1. 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*)

2b2.2. 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)*

A review of the measure specifications was performed, and the measure logic was reviewed, to ensure that the measure logic appropriately captures the clinical intent of the measure. Upon confirmation that the clinical intent was met, the measure logic was loaded into Bonnie, to test the measure logic. The data elements and logic statements confirm that results, based on simulated patients, met expectations in measure performance.

Numerator test cases are added to ensure that patients who did not have a bone scan performed at any time since diagnosis of prostate cancer pass the measure. Denominator test cases are added to ensure that all patients, regardless of age, with a diagnosis of prostate cancer at low risk of recurrence, receiving interstitial prostate brachyotherapy, OR external beam radiotherapy to the prostate , OR radical prostatectomy, OR cryotherapy are included in the measure, as intended. Exception test cases ensure that patients are appropriately removed from the numerator in the case that medical or system exceptions have been identified by the clinician. Examples of medical reason exceptions may include documented pain, salvage therapy or other medical reasons. An example of a system reason exception may include a bone scan being ordered by someone other than the reporting physician.

Face validity of the measure score as an indicator of quality was also systematically assessed, as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

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

Bonnie results provide coverage and passing rates. This measure currently has achieved 100% coverage, which indicates that there is a test case included in the system to evaluate each pathway of the measure logic. The measure also has a 100% passing rate, which indicates that all test cases perform as expected.

Face Validity

The expert panel included 17 members. Panel members were comprised of experts from the PCPI Measures Advisory Committee. The list of expert panel members is as follows:

Joseph Drozda, MD, FACC (Chair)
Richard Bankowitz, MD, MBA, FACP
Heidi Bossley, MSN, MBA
John Easa, MD, FIPP
Christine Goertz, DC, PhD
Jeff Jacobs MD, FACS, FACC, FCCP
Yosef Khan MD, MPH, PhD, MACE
Dianne Jewell, PT, DPT, PhD, FAACVPR
Scott T. MacDonald, MD
Mark Metersky, MD
Michael O'Dell, MD, MS, MSHA, FAAFP
Martha Radford, MD, FACC, FAHA
Amy Sanders, MD, MS
David Seidenwurm, MD
Shannon Sims, MD, PhD
Jessie Sullivan, MD
Karen Johnson (NQF Liaison) - RECUSED

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

The results of the expert panel rating of the validity statement were as follows: N = 10; Mean rating = 3.8 and 80% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)
2 – 2 responses
3 – 0 responses (Neither Agree nor Disagree)
4 – 6 responses
5 – 2 responses (Strongly Agree)

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

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

Exceptions in the eCQM are harmonized with the registry version of this measure (#0389) and include:

- Documentation of reason(s) for performing a bone scan (including documented pain, salvage therapy, other medical reasons, bone scan ordered by someone other than reporting physician)

Exceptions were analyzed for frequency across providers, via the testing project for #0389.

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

Exceptions Analysis results for #0389:

Amongst the 24 physicians with the minimum (10) number of quality reporting events, there were a total of 183 exceptions reported. The average number of exceptions per physician in this sample is 7.6. The overall exception rate is 14.1%.

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

Exceptions are necessary to account for those situations when it is medically appropriate for a patient to have a bone scan. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for several reasons. Rather than specifying an exhaustive list of explicit reasons for exception for this measure, the measure developer relies on clinicians to link the exception with a specific reason for the decision to order a bone scan required for a patient.

Some have indicated concerns with exception reporting including the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234.

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

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. 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

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

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Not applicable

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

Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

Not applicable

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

Not applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

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

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

Measures of central tendency, variability, and dispersion were calculated.

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

Based on the sample of 24 included physicians, included in the analysis for the registry version of this measure (#0389), the mean performance rate is 0.89 the median performance rate is 1.0 and the mode is 1.0. The standard deviation is 0.20. The range of the performance rate is 0.49, with a minimum rate of 0.51 and a maximum rate of 1.00. The interquartile range is 0.12 (0.88 – 1.00).

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

Data 2 (Registry)

The range of performance from 0.51 to 1.00, based on the analysis of the registry version of this measure (#0389), suggests there's clinically meaningful variation across physicians' performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

This test was not performed for this measure.

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

This test was not performed for this measure.

2b6.3. 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*)

This test was not performed for this measure.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

Data are not available to complete this testing.

2b7.2. 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)

Data are not available to complete this testing.

2b7.3. 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*)

Data are not available to complete this testing.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

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

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

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.**

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.**

Attachment **Attachment:** NQF2963_FeasibilityAssessBonnieAttachment.pdf

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. **Describe what you have learned/modified 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 a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA). Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

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.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting PQRS http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/pqrs/index.html Payment Program Meaningful Use Stage II https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/

4a.1. For each CURRENT use, checked above, provide:

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

1.Physician Quality Reporting System (PQRS) – Sponsored by the Centers for Medicare and Medicaid Services (CMS)

Purpose: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013. Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented a phased approach to public reporting performance information on the Physician Compare Web site. CMS announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in late 2017.

2.Meaningful Use Stage 2 (EHR Incentive Program) – Sponsored by the Centers for Medicare and Medicaid Services (CMS)

Purpose: The Medicare and Medicaid EHR Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology.

Eligibility for incentive payments for the "meaningful use" of certified EHR technology is established if all program requirements are met, including successful implementation and reporting of program measures, which include this measure, to demonstrate meaningful use of EHR technology.

4a.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?*)

4a.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.*)

4b. 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.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Report Title: PQRS Ad Hoc Analysis PQ3394, 2014 PQRS Measure Data for PCPI
Report includes Final Action 2014 EHR data, Final Action 2014 Registry Data and Part B Claims data for services rendered between January 1, 2014 and December 31, 2014 and processed into NCH by February 27, 2015.

01/01/2014 – 12/31/2014 EHR Performance Rate:
Mean: 90.76%
Maximum: 100.00%
Minimum: 50.00%

01/01/2014 – 12/31/2014 Registry Performance Rate:
Mean: 90.24%
Minimum: 0.00%
Maximum: 100.00%

2013 PQRS Experience Report by Individual Measure:
2013 is the most recent year for which PQRS Experience Report measure data is available. The average performance rates on over Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients the last several years are as follows:
2010: 71.60%
2011: 90.50%
2012: 92.50%
2013: 88.50%

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2013, 8.2% of eligible professionals reported on Avoidance of Overuse of Bone Scan for Staging Low Risk Prostate Cancer Patients for claims, registry, and electronic health records. As a result, performance rates may not be nationally representative.
Reference: Center for Medicare and Medicaid Services. 2013 Reporting Experience Including Trends. Available:
https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

4b.2. **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.**
While the PCPI creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1.Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun

5;309(21):2215-6.

4c. 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).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

We are not aware of any unintended consequences at this time, but we take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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)**
0390 : Prostate Cancer: Adjuvant Hormonal Therapy for High or Very High Risk Prostate Cancer Patients
1853 : Radical Prostatectomy Pathology Reporting

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

5a. Harmonization

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 completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The related measure 1853, Radical Prostatectomy Pathology Reporting, addresses the percentage of radical prostatectomy pathology reports that include the pT category, the pN category, the Gleason score and a statement about margin status, which is a different action than measure 0389. The two measures do not share similar target populations and address different aspects of prostate cancer care. The related measure 0390, Prostate Cancer: Adjuvant Hormonal Therapy for High Risk or Very High Risk Prostate Cancer Patients addresses the use of adjuvant hormonal therapy and external beam radiation therapy in high-risk prostate cancer patients which is a different quality action from measure 0389. The two measures do not share similar target populations and address different aspects of prostate cancer care.

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.
No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** PCPI
Co.2 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-
Co.3 **Measure Developer if different from Measure Steward:** PCPI
Co.4 **Point of Contact:** Diedra, Gray, diedra.gray@ama-assn.org, 312-464-4904-

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.
Ian Thompson, MD (Co-Chair, urology)
Steven Clauser, PhD (Co-Chair, methodology)
Peter Albertsen, MD (urology)
Colleen Lawton, MD (radiation oncology)
Charles Bennett, MD, PhD, MPP (clinical oncology)
W. Robert Lee, MD, MS, Med (radiation oncology)
Michael Cookson, MD (urology)
Peter A. S. Johnstone, MD, FACR (radiation oncology)
Gregory W. Cotter, MD (radiation oncology)
David F. Penson, MD, MPH (urology)
Theodore L. DeWeese, MD (radiation oncology)
Stephen Permut, MD (family medicine)
Mario Gonzalez, MD (pathology)
Howard Sandler, MD (radiation oncology)
Louis Kavoussi, MD (urology)
Bill Steirman, MA (consumer representative)
Eric A. Klein, MD (urology)
John T. Wei, MD (urology)
Carol Wilhoit, MD (health plan representative)

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 09, 2015
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, 2017

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

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. The AMA, the PCPI and its members and former members of the AMA-PCPI disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:** Please see the copyright statement above in AD.6 for disclaimer information.

Ad.8 **Additional Information/Comments:**