**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0220

**Measure Title**: Adjuvant hormonal therapy is recommended or administered within 1 year (365 days) of diagnosis for women with AJCC T1cN0M0 or Stage IB – Stage III hormone receptor positive breast cancer

**Date of Submission**: August 1, 2019

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: 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 nationally recognized National Cancer Database (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 approximately 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 80 percent of newly diagnosed breast cancer cases nationwide.

The NCDB collects data from CoC-accredited cancer programs on an annual basis; the data collected is in accordance with standard registry procedures. In January of 2019, 2016 diagnoses were processed and included in the data warehouse. As of this submission, 2017 diagnoses were not yet processed. We find information on some of the therapies which take longer to receive 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 performance rates 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 data year until programs have had time to submit treatment data and as such 2016 data were not included. Throughout this testing document, there are three patient cohorts presented, which represent combined 2014 and 2015 diagnoses, and 2014 and 2015 separately.

**1.3. What are the dates of the data used in testing**? Click here to enter date range

January 1, 2014-December 31, 2015

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

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

Testing for measure #0220 was performed with 1,426 Commission on Cancer-accredited hospitals across the United States from 2014-2015.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnosis Year(s) | | 2014 - 2015 | 2014 | 2015 |
| Hospital N | | 1,426 | 1,410 | 1,360 |
| Range of Cases per Hospital | | 1 - 1,939 | 1 - 1,025 | 1 - 914 |
| Hospital Category Type | Academic | 189 (13.25%) | 186 (13.19%) | 188 (13.82%) |
| Community | 397 (27.84%) | 393 (27.87%) | 388 (28.53%) |
| Comprehensive Community | 571 (40.04%) | 565 (40.07%) | 561 (41.25%) |
| Integrated Network | 155 (10.87%) | 154 (10.92%) | 153 (11.25%) |
| NCI & PPS-Exempt | 45 (3.16%) | 45 (3.19%) | 44 (3.24%) |
| Other | 69 (4.84%) | 67 (4.75%) | 26 (1.91%) |

**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*)

The 2014-2015 testing data included 164,893 cases, all of which are measure-eligible.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnosis Year(s) | | 2014 - 2015 | 2014 | 2015 |
| Case N | | 164,893 | 83,253 | 81,640 |
| Race/ Ethnicity | Non-Hispanic White | 125,759 (76.27%) | 63,498 (76.27%) | 62,261 (76.26%) |
| Non-Hispanic Black | 16,502 (10.01%) | 8,354 (10.03%) | 8,148 (9.98%) |
| Asian, Hawaiian, Pacific Islander | 5,600 (3.40%) | 2,774 (3.33%) | 2,826 (3.46%) |
| Hispanic | 10,130 (6.14%) | 5,079 (6.10%) | 5,051 (6.19%) |
| Other | 6,902 (4.19%) | 3,548 (4.26%) | 3,354 (4.11%) |
| Age at Diagnosis | 18 - 49 | 37,179 (22.55%) | 18,771 (22.55%) | 18,408 (22.55%) |
| 50 - 59 | 42,059 (25.51%) | 21,584 (25.93%) | 20,475 (25.08%) |
| 60 - 69 | 46,557 (28.23%) | 23,333 (28.03%) | 23,224 (28.45%) |
| 70 - 79 | 28,969 (17.57%) | 14,337 (17.22%) | 14,632 (17.92%) |
| 80 + | 10,129 (6.14%) | 5,228 (6.28%) | 4,901 (6.00%) |
| Insurance Status | Not Insured, Medicaid | 12,957 (7.86%) | 6,646 (7.98%) | 6,311 (7.73%) |
| Private | 90,260 (54.74%) | 45,711 (54.91%) | 44,549 (54.57%) |
| Medicare | 57,502 (34.87%) | 28,761 (34.55%) | 28,741 (35.20%) |
| Other Government | 1,841 (1.12%) | 932 (1.12%) | 909 (1.11%) |
| Other/Unknown | 2,333 (1.41%) | 1,203 (1.44%) | 1,130 (1.38%) |

**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**.

Not Applicable

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

Not Applicable

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

**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*)

There are three characteristics that generally inform the reliability of a measure including hospital-level performance, the variability of performance differences across hospitals, and case volume. In an effort to illustrate the impact of performance and volume, reliability is presented in three ways. Three patient cohorts constructed represent combined 2014 and 2015 diagnoses, and 2014 and 2015 separately. Reliability results are shown for each cohort. Within each cohort, data model reliability estimations are shown computed from both unadjusted and risk-adjusted models. Risk-adjustment included patient characteristics of age at diagnosis, race/ethnicity, and insurance status.

Measure compliance was modeled from 2-level hierarchical logistic regression models using Bayesian shrinkage adjustments that control for random error for both patients and hospitals. Statistical reliability is determined with a binary-outcome from two types of variability, between hospitals (signal) obtained from the regression model and within hospitals (noise) based on the standard error of the proportion of the hospital random effect. Reliability is presented here on a scale from 0 to 1 from a range indicating measurement error to true differences in hospital performance. Statistical reliability was converted from the log-odds scale to the probability scale through the hierarchical method of calculation as described by Deutsch et al. and referenced in other publications 1 2 3 4 5.

1Deutsch A, Smith L, Gage B, et al. Patient-reported outcomes in performance measurement: commissioned paper on PRO-based performance measures for healthcare accountable entities draft no. 1. Prepared for NQF by RTI International and the Brookings Institution. September 4, 2012. Available at: https://www.qualityforum.org/Projects/n-r/Patient-Reported\_Outcomes/ Commissioned\_Paper\_2.aspx.

Accessed on September 28, 2017.

2Liu JB, Huffman KM, Palis BE, et al. Reliability of the American College of Surgeons Commission on Cancer’s quality of care measures for hospital and surgeon profiling. *Journal of the American College of Surgeons*. 2017; 224: 180-190.

3Lawson EH, Ko CY, Adams JL, et al. Reliability of evaluating hospital quality by colorectal surgical site infection type. *Annuls of Surgery*. 2013; 258: 994–1000.

4Huffman KM, Cohen ME, Ko CY, Hall BL. A comprehensive evaluation of statistical reliability in ACS NSQIP profiling models. *Annuls of Surgery*. 2015; 261:1108–1113.

5Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *Journal of the American College of Surgeons.* 2013; 217: 336-346.

**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*)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Diagnosis Year(s) | Denominator | Numerator | Hospital N | Unadjusted Reliability Mean | Unadjusted Reliability Median | Adjusted Reliability Mean1 | Adjusted Reliability Median1 | Mean Cases (per hospital) | Aggregate Compliance  Rate |
| 2014 - 2015 | 164,893 | 152,733 | 1,426 | 0.83 | 0.90 | 0.83 | 0.90 | 115.63 | 92.63% |
| 2014 | 83,253 | 77,056 | 1,410 | 0.74 | 0.81 | 0.73 | 0.80 | 59.04 | 92.56% |
| 2015 | 81,640 | 75,677 | 1,360 | 0.77 | 0.83 | 0.76 | 0.83 | 60.03 | 92.70% |
| 1 Adjusted for age at diagnosis, race/ethnicity and insurance status | | | | | | | | | |

**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 unadjusted reliability mean coefficient of 0.74, 0.77, and 0.83 for measure #0220 are all regarded as very good, achieved respectively from diagnosis years 2014, 2015, and 2014-2015.

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

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

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

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

Each year, a review of a minimum of 10% of the annual caseload of registry abstracts is performed to verify that abstracted data correctly reflect the information documented in individual patient records, and that the patient’s medical condition, care, and participation in treatment decision-making processes are accurate. The abstracted elements reviewed include but are not limited to primary site, staging, first-course treatment, follow-up, and the percentage of data coded as unknown. These procedures are part of the Commission on Cancer’s (CoC) Standard 1.6, which is required to maintain accreditation. Each of the following steps must be followed in order to be rated as compliant with this Standard. 1) The cancer committee establishes and implements a quality control plan. 2) The registry quality coordinator works cooperatively with registry staff to maintain a quality control plan. The focus of this plan is to establish data quality benchmarks that include monitoring of abstracting timeliness, accuracy of data, a review of data coded as unknown. 3) The findings are to be reported to the cancer committee annually and 4) the findings are documented in the cancer committee minutes.

These annual caseload reviews inform the data submitted to the National Cancer Database (NCDB) and are subsequently used in the Cancer Program Practice Profile Reports (CP3R) system and Rapid Quality Reporting System (RQRS) measure reporting systems; RQRS shows the performance rates for the five CoC National Quality Forum endorsed measures. Every breast cancer case is submitted from the reporting facility to the NCDB, and is applied to the measure in these reporting systems. In addition to the annual caseload reviews noted above, both reporting systems allow hospitals to review every case for coding accuracy which includes those deemed non-eligible, incomplete, measure denominator eligible, along with numerator compliant and non-compliant cases.

Measure performance is calculated in CP3R and RQRS based on the case-level data submitted by the accredited hospitals.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
There are no statistical tests run to assess quality of the data. Non-compliance results in any one or more of the four steps listed in 2b1.2 (Standard 1.6) not followed will potentially jeopardize the accreditation status of the program granted by the Commission on Cancer (CoC). The authoritative source for comparison is the patient chart with the goal of this standard to ensure the registry abstract reflects the documented patient experience. The scope of the evaluation is 10% of the analytic caseload for each program or a maximum of 300 cases annually.

The measure reporting system reviews are highly recommended by the CoC to ensure high quality data, which directly impacts the performance rates. The Cancer Program Practice Profile Reports system performance rate for this measure is directly tied to CoC Standard 4.4, which maintains that a rate of 90% is met or exceeded, or that the program has implemented an action plan that reviews and addresses program performance below the expected Estimated Performance Rates. Failure to follow these directives results in non-compliance with Standard 4.4 which, like Standard 1.6, can potentially jeopardize the accreditation status of the program granted by the CoC.

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)  
Both the annual caseload reviews and the measure reporting system reviews are intended to ensure that reported performance rates are an accurate reflection of the care provided to patients at Commission on Cancer-accredited programs.

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

**2b2. EXCLUSIONS ANALYSIS**

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

Not Applicable

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

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)  
Not Applicable

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)  
Not Applicable

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

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

Not Applicable

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Not Applicable

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Not Applicable

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

Not Applicable

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**Not Applicable

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

Not Applicable

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

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

Not Applicable

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**Not Applicable

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not Applicable

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Not Applicable

**2b3.9. Results of Risk Stratification Analysis**:

Not Applicable

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
Not Applicable

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

Not Applicable

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
Ninety-five percent confidence intervals were calculated at the hospital-level to determine statistical significance. Each hospital’s 95% confidence interval was compared to the aggregate EPR to determine significance.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
The following table calculates for each set of diagnosis year(s) the aggregate measure compliance and the number of hospitals that were not statistically significant compared to the aggregate measure compliance and the hospitals that were statistically significant (both greater and less than aggregate measure compliance):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnosis Year(s) | Aggregate Measure Compliance | N Hospitals Statistical Significance Compared to Aggregate Measure Compliance | | |
| Not Significant | Significantly Greater Than | Significantly Less Than |
| 2014 - 2015 | 92.63% | 716 (50.21%) | 548 (38.43%) | 162 (11.36%) |
| 2014 | 92.56% | 769 (54.54%) | 520 (36.88%) | 121 (8.58%) |
| 2015 | 92.70% | 685 (50.37%) | 554 (40.74%) | 121 (8.90%) |

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
For the three cohorts of diagnosis years, slightly over half of the programs were not statistically significant compared to aggregate measure compliance; however for those statistically significant programs the majority were statistically greater than the aggregate measure compliance. Overall, only between 8.58% and 11.36% of the programs was performing worse when compared to the aggregate.

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

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

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

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

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

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
Not Applicable

**2b5.3. What is your interpretation of the results in terms of 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

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)  
All cancer cases submitted from Commission on Cancer-accredited programs are evaluated for measure eligibility via the Cancer Program Practice Profile Reports system and Rapid Quality Reporting System. Both systems track the completeness of staging and all data fields specifically needed to assess measure #0220 denominator including tumor size, hormone receptor status, and date completeness. For each hospital, cases with incomplete data were summed and reported as a proportion of all denominator-eligible and incomplete cases.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnosis Year(s) | | 2014 - 2015 | 2014 | 2015 |
| Measure | Aggregate Compliance | 92.63% | 92.56% | 92.70% |
| Aggregate Denominator | 164,893 | 83,253 | 81,640 |
| Aggregate Hospital N | 1,426 | 1,410 | 1,360 |
| Incomplete and Complete Data | Aggregate Incomplete | 0.09% | 0.09% | 0.09% |
| Aggregate Denominator | 165,046 | 83,332 | 81,714 |
| No Incomplete Data for N Hospitals | 1,345 | 1,369 | 1,312 |
| Incomplete Data for N Hospitals | 81 | 41 | 48 |
| Range of Incomplete Data | 0.12% - 46.43% | 0.38% - 46.43% | 0.24% - 16.67% |

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

For 2014-2015 data completeness varied at the hospital-level resulting in a range of 0.12% to 46.43% for the 5.68% of hospitals that had incomplete cases. At the aggregate-level, the proportion of incomplete cases was low, 0.09%, when compared to the number of measure-eligible and incomplete cases. While missing data will always bias compliance rates, we do not believe the proportion to be excessive and therefore not detrimental to the measure. Given that the Commission on Cancer mandates all cancer cases be submitted from a reporting hospital and that all submissions assessed for measure compliance, we expect and allow varying degrees of reporting incompleteness. The measure reporting systems allow hospitals to review case-level data for incomplete cases and are encouraged to make completeness updates.