**NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0507

**Measure Title**: Diagnostic Imaging: Stenosis Measurement in Carotid Imaging Reports

**Date of Submission**: 1/1/2021  
**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 specifications** (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**](#_bookmark0) 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**](#_bookmark1) 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**](#_bookmark2)

# 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**](#_bookmark3)

**2b3. For outcome measures and other measures when indicated** (e.g., resource use):

* **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care;[**14,15**](#_bookmark4) 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**](#_bookmark5) **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**

1. 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).
2. 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.
3. 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.
4. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
5. Risk factors that influence outcomes should not be specified as exclusions.
6. 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.

# 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. **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. **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 American College of Radiology (ACR) completed measure testing using Medicare Part B claims, qualified registry data, and qualified clinical data registry data. The data was obtained from the Centers for Medicare & Medicaid Services (CMS).

* + 1. **What are the dates of the data used in testing**?
* The data collection period was from 2010
* The data collection period was 2012-2014

The most recent measure testing data is from January 1, 2015 - December 31, 2018

* + 1. **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. **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*)

Initial Testing Project (2010):

* Three radiology practice sites representing various types, locations and sizes were identified to participate in testing the measures
* The number of physicians per site was between 10 and 1,000 physicians
* Two of the sites were hospital-based radiology group practices and one was a stand-alone radiology group practice
* All three sites were located in urban regions
* Patient visit volume ranged from 550-1600 patients, per site, per day
* Sample size included a total of 109 records for this measure
* The data collection period was 1/1/2010- 12/31/2010
* Data abstraction was performed in 2011

(2015)

The numbers of physicians were 133,717 physicians

Among these physicians 128, 525 data was based on claims and 5192 was based on registry data

**(2020)**

The testing sample comprised all NPIs who submitted data to CMS for this measure. The sample consisted of 55,761 physicians. The eligible population for this measure (i.e. the denominator) includes all final reports for carotid imaging studies (neck MR angiography [MRA], neck CT angiography [CTA], neck duplex ultrasound, carotid angiogram) performed. There are no exclusions to this measure.

**Table 1. Number of providers that submitted data for this measure.**

|  |  |
| --- | --- |
|  | **# of NPIs** |
| ***All- Claims, QCDR, and Registry*** | |
| ***All 4 Years*** | 55,761 |
| ***2015*** | 15,095 |
| ***2016*** | 17,722 |
| ***2017*** | 12,713 |
| ***2018*** | 10,231 |
| **Claims** | |
| ***All 4 Years*** | 47,893 |
| ***2015*** | 12,729 |
| ***2016*** | 14,201 |
| ***2017*** | 11,759 |
| ***2018*** | 9,204 |
| **QCDR** | |
| ***Both Years \**** | 365 |
| ***2015*** | 215 |
| ***2016*** | 150 |
| **Registry** | |
| ***All 4 Years*** | 7,503 |
| ***2015*** | 2,151 |
| ***2016*** | 3,371 |
| ***2017*** | 954 |
| ***2018*** | 1,027 |

*\*CMS combined QCDR and Registry data beginning in 2017.*

* + 1. **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*)

**(2015)**

• Number of patients eligible were 6,877,159 (avg. per NPI is 51.43)

• Number of patients reported were 2,268,250 (avg. per NPI is 16.96)

**(2020)**

A total of 7,267,917 individuals were eligible to be included in this testing. However, for the Merit-based Incentives Payment System (MIPS), physicians are not required to submit all patient data to CMS. Between 2016 and 2018, a minimum 60% of data was required for reporting. The ACR performed testing with the 6,462,722 individuals that were reported to CMS.

**Table 2. Eligible Patients and Reported Patients**

|  |  |  |
| --- | --- | --- |
|  | **# of Patients Eligible** | **# of Patients Reported** |
| ***All- Claims, QCDR, and Registry*** | |  |
| ***All 4 Years*** | 7,267,917 | 6,462,722 |
| ***2015*** | 1,123,433 | 982,806 |
| ***2016*** | 1,550,485 | 1,387,545 |
| ***2017*** | 1,887,183 | 1,627,953 |
| ***2018*** | 2,706,816 | 2,464,418 |
| **Claims** | |  |
| ***All 4 Years*** | 3214500 | 2,577,610 |
| ***2015*** | 860,528 | 724,920 |
| ***2016*** | 949,197 | 802,753 |
| ***2017*** | 740,179 | 564,876 |
| ***2018*** | 664,596 | 485,061 |
| **QCDR** | |  |
| ***Both Years\**** | 269,815 | 257,627 |
| ***2015*** | 26,114 | 25,225 |
| ***2016*** | 243,701 | 232,402 |
| **Registry** | |  |
| ***All 4 Years*** | 3,783,602 | 3,627,485 |
| ***2015*** | 236,791 | 232,661 |
| ***2016*** | 357,587 | 352,390 |
| ***2017*** | 1,147,004 | 1,063,077 |
| ***2018*** | 2,042,220 | 1,979,357 |

*\* CMS combined QCDR and registry data beginning in 2017.*

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

There are no differences in the data or sample used for different aspects of testing.

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

No social risk factors were analyzed 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*)

ACR performed a signal-to-noise ratio (SNR) analysis test on the performance data for reliability. In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers.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.

A reliability score equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability score equal to one implies that all the variability is attributable to real differences in physician performance. A reliability score of 0.70 is generally considered the minimum threshold for reliability and 0.80 is generally considered very good reliability.

SNR reliability testing is performed using the Beta-Binomial Model, which assumes that physicians’ performance scores are a binomial random variable conditional on the physicians’ true value derived from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta are considered intermediate calculations used to establish the variance estimates.

ACR testing protocol 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. The minimum threshold of events was set at 10. Limiting the reliability analysis to physicians with a minimum number of events reduces bias introduced by the inclusion of physicians without a significant number of events.

CMS physician-level claims, registry, and QCDR data was extracted for the relevant physician-level information.

**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*)

**(2015)**

The following are the results of inter-rater reliability testing.

Reliability (N, % Agreement, Kappa ( 95% Confidence Interval))

Overall Reliability (109, 100%, n/a\*)

Numerator Reliability (109, 100%, n/a\*)

Denominator Reliability (109, 100%, n/a\*)

\* 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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Summary of PQRS Reliability Score Stats Cumulative and by Year (2012 - 2014)*** | | | | | | |  |
|  |  |  |  |  |  |  |  |
| *Year* | *Number of Providers* | *Reliability p25* | *Reliability median* | *Reliability p75* | *Reliability mean* | *Reliability LCLM* | *Reliability UCLM* |
| 2012 | 37142 | 0.81728 | 1 | 1 | 0.84524 | 0.84238 | 0.84809 |
| 2013 | 33493 | 0.63205 | 1 | 1 | 0.81781 | 0.81477 | 0.82085 |
| 2014 | 12953 | 0.99306 | 0.99722 | 0.99913 | 0.99473 | 0.99461 | 0.99484 |
| All | 83588 | 0.88581 | 1 | 1 | 0.85741 | 0.85561 | 0.85922 |

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 0.85741 (0.85561, 0.85922). A reliability of 0.80 is considered very good reliability. So according to the reliability testing analysis, the results demonstrated very good reliability.

**(2020)**

Using the parameter estimates from the beta-binomial model, we computed reliability scores for each performance year. Please see **Table 3** for the results.

**Table 3. Reliability Score Statistics by Year by Provider (claims and registry)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Year*** | ***Number of Providers*** | ***25th percentile*** | ***Reliability median*** | ***75th percentile*** | ***Reliability mean*** | ***Lower Confidence Limit (minimum)*** | ***Upper Confidence Limit (maximum)*** |
| 2015 | 15095 | .99467 | .99804 | .99950 | .99593 | .99584 | .99602 |
| 2016 | 17994 | .99576 | .99857 | .99983 | .99680 | .99674 | .99687 |
| 2017 | 13579 | .98857 | .99601 | .99957 | .99185 | .99167 | .99203 |
| 2018 | 11133 | .98352 | .99481 | .99960 | .98638 | .98599 | .98677 |
| **ALL** | 57801 | .99247 | .99765 | .99967 | .99340 | .99331 | .99350 |

**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 0.99308 (0.99303, 0.99313). A reliability of 0.80 is considered very good reliability. Therefore, according to the reliability testing analysis, the results demonstrate very good reliability.

2020 Update:

This measure remains consistently reliable. The mean (CI) reliability is 0.99340 (0.99331, 0.99350), which is higher than the required 0.80.

# 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)*

Our expert panel included 14 members whose specialties include neuroradiology, abdominal radiology, musculoskeletal radiology, cardiac/thoracic radiology, breast imaging, general diagnostic radiology, nuclear medicine, informatics, quality, and physics.

* David Seidenwurm, MD, FACR (Chair) (Radiology/Neuroradiology) Sacramento, CA
* Dorothy Bulas, MD, FACR (Radiology/Pediatric Radiology) Washington, DC
* Robert Henkin, MD, FACR (Nuclear Medicine)
* Charles Johnson, MD, FACR (Radiology/Abdominal Radiology) Scottsdale, AZ
* David Rubin, MD (Radiology, Musculoskeletal Radiology) Saint Louis, MO
* Frank Rybicki, MD (Radiology/Cardiac/thoracic Radiology) Boston, MA
* Elizabeth Burnside, MD, MPH (Radiology/Breast Imaging) Madison, WI
* Matt Hawkins (Radiology Fellow) Cincinnati, OH
* Jonathan Kruskal, MBChB, PhD (Radiology/Abdominal Radiology) Newton, MA
* Frank Lexa, MD, MBA (Radiology/Neuroradiology) Wynnewood, PA
* Paul Nagy, PhD (Informatics, Quality, Physicist) Baltimore, MD
* Donald Renfrew, MD (General Diagnostic Radiology) Sturgeon Bay, WI
* Bob Pyatt, MD (General Diagnostic Radiology) Chambersburg, PA
* Paul Larson, MD (General Diagnostic Radiology) Madison, WI

This performance measure was assessed for content validity by a panel of 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 specifically for this purpose. All comments received are reviewed by the expert work group and the measures adjusted as needed. Other external review groups (eg, focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

An expert panel was used to assess face validity of the measure. This panel consisted of 14 members, with representation from the following specialties: neuroradiology, abdominal radiology, musculoskeletal radiology, cardiac/thoracic radiology, breast imaging, general diagnostic radiology, nuclear medicine, informatics, quality, and physics.

The aforementioned panel was asked to rate their agreement with the following statement:

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

2020 Update:

ACR endeavored to perform construct validity empirical testing on NQF # 0507. This measure requires radiologists use a standardized, validated method for quantifying stenosis in carotid imaging studies, specifically direct or indirect measurements of the distal internal carotid diameter. The measure purpose is to improve reporting the method used to assess the degree of stenosis, Since degree is a critical factor in determining the treatment and management approach for patients with carotid stenosis, how the degree of stenosis is assessed should be standardized. By comparing NQF # 0507 performance data to data of a related measure, we intended to hypothesize that the performance of the related measure correlated with the performance of NQF #507. However, we were unable to identify a measure suitable for comparison within the same accountability program (MIPS) for which we could obtain *patient* level data. We used the CMS Measure Repository to search for related measures.

ACR identified two measures in the MIPS program that are related to #0507: MIPS #413, *Door to Puncture Time for Endovascular Stroke Treatment*, MIPS #409, *Clinical Outcome Post Endovascular Stroke Treatment*. We also identified a related measure in the CMS Hospital Outpatient Quality Reporting (HOQR) program, NQF #0661, OP-23: *Head CT scan results for acute ischemic stroke or hemorrhagic stroke who received head CT scan interpretation within 45 minutes of arrival*. Each of these measures focus on timely and effective care for stroke care and treatment. For these measures, we were only able to obtain *population-*level measure data. NQF indicated that an acceptable alternate option for demonstrating empirical validity is to perform criterion validity using measure performance data at the *population* level. Our plan to perform the requisite analyses among these measures to determine if a relationship exists to support empirical validity, hypothesizing that hospitals or physicians performing well on these measures (MIPS #409, MIPS #413 and HOQR OP-23/NQF #0661) would perform the same on the stenosis measure (NQF #0507). However, we were unable to format the measures’ data sets to perform empirical analysis. While MIPS #409 and #413 are specified at the individual clinician level, CMS was unable to provide us with individual level data, because all submissions were done at the group level. In addition, there were 26 eligible submissions between 2016 and 2018 that had performance rates. The specialties that submitted data for these measures were non-radiologists. Given the shortage of data, the group level data submission, and the different clinician types, the performance on NQF #0507 could not be compared with MIPS #409 and MIPS #413. The data from HOQR OP-23/NQF #0061 was missing the number of clinicians that performed the measure, the patient sample sizes, the numerator and denominator, so it was also unable to be correlated with NQF #0507.

Due to the lack of appropriate measurement data to perform empirical validity testing, ACR performed a new face validity survey on this measure in November 2020. An expert panel assessed the face validity of the measure. The panel consisted of 28 members in a cross-section of practice types and geographical locations.

* Amy Kotsenas, MD (Neuroradiology) Rochester, MN
* Rajeev Shah, MD, MBA (Diagnostic Radiology/Neuroradiology) Austin, TX
* Cathrine Keller, MD (Diagnostic Radiology) Leesburg, FL
* Brian Berger, MD (Diagnostic Radiology) Shelby Township, MI
* Yvonne Lui, MD (Neuroradiology/AI) New York, NY
* Ajay Gupta, MD (Diagnostic Radiology/Neuroradiology) New York, NY
* Haris Sair, MD (Diagnostic Radiology/AI) Baltimore, MD
* Jeffrey Jarvik, MD, MPH (Neuroradiology) Seattle, WA
* Max Wintermark, MD (Neuroradiology) Stanford, CA
* Michael Iv, MD (Neuroradiology) Palo Alto, CA
* Jeffrey Stone, MD (Diagnostic Radiology) Jacksonville, FL
* Theodore Larson III, MD (Interventional Radiology/Neuroradiology) Centennial, CO
* Steven Falcone, MD (Neuroradiology) Miami, FL
* Gloria Guzman, MD, MSc, MPH (Neuroradiology) St. Louis, MO
* Jody Tanabe, MD (Diagnostic Radiology/Neuroradiology) Aurora, CO
* Achala Vagal, MD (Neuroradiology) Cincinnati, OH
* Sammy Chu, MD (Diagnostic Radiology) Bellevue, WA
* Nolan Kagetsu, MD (Diagnostic Radiology) New York, NY
* Bradley Delman, MD (Neuroradiology) New York, NY
* John Jordan, MD (Neuroradiology) Torrance, CA
* Fabio Settecase, MD (Diagnostic Radiology) San Francisco, CA
* Patrick Turski, MD (Diagnostic Radiology/Medical Physics) Madison, WI
* Mariya Gusman (Neuroradiology Fellow) Fairfield, CA
* Jacob Ormsby, MD, MBA (Diagnostic Radiology/Neuroradiology) Albuquerque, NM
* William Donovan, MD, MPH (Neuroradiology) Norwich, CT
* Noushin Yahyavi Firouz Abadi, MD (Neuroradiology) Potomac, MD
* Roland Lee, MD (Neuroradiology) San Diego, CA
* Salil Soman, MD (Diagnostic Radiology/Neuroradiology) Boston MA

The panel was asked to rate their agreement with the following statement: The scores obtained from the measure, as specified, will accurately differentiate quality across providers. The panel could choose from a scale of 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree.

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

The results of the expert panel rating of the validity statement were as follows: N = 7; Mean rating = 4.43 and 85.71% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

This measure underwent maintenance review by an expert panel. The review was completed in February 2015. New evidence was reviewed. The expert panel agreed that the measure remained valid based on existing and new evidence.

82.15% (23 members) of the panel either strongly agreed or agreed that this measure accurately distinguishes good from poor quality. Two panel members disagreed that the measure would accurately distinguish good from poor quality. One member stated that literature shows that CTA underestimates the stenosis and MRA overestimates the stenosis, compared to NASCET. The other member stated that ultrasound should be removed from the measure. He also added that stenosis on ultrasound is measured using velocities and NASCET should never be applied to ultrasound. However, the ACR respectfully disagrees and believes perhaps the commenter may not have fully understood the goal or construct of the measure. The types of imaging included in the measure (MRA, CTA, duplex US, angiography) are all used in practice to evaluate carotid artery stenosis. The intent of the measure is to improve consistency in reporting the method used to estimate stenosis, agnostic to the modality/technology of the carotid imaging study used to evaluate level of stenosis. Evidence shows that NASCET has standardized the method of quantifying stenosis, specifically with reference to the distal internal carotid diameter as the denominator for stenosis measurement, in several different types of imaging studies. There are validated methods to cross reference and correlate indirect reference to the carotid diameter to angiography or other imaging studies, which are gold standard. For example, in Doppler ultrasound, the degree of stenosis can be estimated using Doppler parameter of the peak systolic velocity (PSV) of the internal carotid artery (ICA), with concordance of the degree of narrowing of the ICA lumen. Additional Doppler parameters of ICA-to-common carotid artery (CCA) PSV ratio and ICA end-diastolic velocity (EDV) can be used when degree of stenosis is uncertain from ICA PSV. (Grant et al, Society of Radiologists in Ultrasound, 2003).

**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?*)

This measure remains valid.

# 2b2. EXCLUSIONS ANALYSIS

**NA** ☒ **no exclusions — *skip to section*** [***2b3***](#_bookmark6)

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

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

# 2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

## If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section [2b4](#_bookmark8).

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

* **No risk adjustment or stratification**
* **Statistical risk model with** Click here to enter number of factors **risk factors**
* **Stratification by** Click here to enter number of categories **risk categories**
* **Other,** Click here to enter description

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

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

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

**Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

* + **Published literature**
  + **Internal data analysis**
  + **Other (please describe)**

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

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

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

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

## If stratified, skip to [2b3.9](#_bookmark7)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**: 2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*): **2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**: **2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified**

(*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat*

*the information provided related to performance gap in 1b)*

To assess statistically significant differences in measure rates, the data described in sections above were used to calculate the mean, median, standard deviation, and interquartile range for the measure rates. In addition, the rates were divided into quartiles, and a Student’s t-test was used to compare the rates of the plans in the 25th percentile to the rates of the plans in the 75th percentile.

**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 tables below show the distribution of measure rates for **claims** data between 2015 and 2018. The mean rate was 94.52%, with a median rate of 100%, minimum rate of 1.27%, and maximum rate of 100%.

**Table 4. Variation in Measure Rates for Claims Data – 2015 to 2018**

|  |  |  |
| --- | --- | --- |
| **Mean** | **Median** | **Standard Deviation** |
| 94.52% | 100% | 14.17% |

**Table 5. Distribution of Measure Rates for Claims Data – 2015 to 2018**

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Minimum | 1.27% |
| 25th percentile | 97.22% |
| 50th percentile (median) | 100% |
| 75th percentile | 100% |
| Maximum | 100% |
| Interquartile Range | 2.78% |
| Student’s t-test p-value | P<0.0001 |

The tables below show the distribution of measure rates for **Registry** data between 2015 and 2018. The mean rate was 98.48%, with a median rate of 100%, minimum rate of 0.06%, and maximum rate of 100%.

**Table 6. Variation in Measure Rates for Registry Data – 2015 to 2018**

|  |  |  |
| --- | --- | --- |
| **Mean** | **Median** | **Standard Deviation** |
| 98.48% | 100% | 6.94% |

**Table 7. Distribution of Measure Rates for Registry Data – 2015 to 2018**

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Minimum | 0.06% |
| 25th percentile | 100% |
| 50th percentile (median) | 100% |
| 75th percentile | 100% |
| Maximum | 100% |
| Interquartile Range | 0 % |
| Student’s t-test p-value | P<0.0001 |

The tables below show the distribution of measure rates for **QCDR** data between 2015 and 2016. As a reminder, the QCDR data for 2017 and 2018 is combined in the registry data above. The mean rate was 293.45%, with a median rate of 100 %, minimum rate of 0%, and maximum rate of 100%.

**Table 8. Variation in Measure Rates for QCDR Data – 2015 to 2016**

|  |  |  |
| --- | --- | --- |
| **Mean** | **Median** | **Standard Deviation** |
| 97.62% | 100% | 8.35% |

**Table 9. Distribution of Measure Rates for QCDR Data – 2015 to 2016**

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Minimum | 18.18% |
| 25th percentile | 100% |
| 50th percentile (median) | 100% |
| 75th percentile | 100% |
| Maximum | 100% |
| Interquartile Range | 0% |
| Student’s t-test p-value | P<0.0001 |

**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 all sets of data data, the measure rates did not show significant variation in the interquartile ranges, but did show a statistically significant difference in the measure rates between the top and bottom quartile of the plans included in the testing (P<0.0001 at alpha = 0.05).

|  |  |
| --- | --- |
| **Submission Method** | **Interquartile Range** |
| Claims | 2.78% |
| Registry | 0% |
| QCDR | 0% |

However, while the variation in the data set appears low, the ACR reviewed the number of eligible instances where physicians could have submitted this measure against the number of instances reported to CMS. The performance rate is high among physicians who chose to report this measure to CMS. However high-performing measures may have low adoption rates among all physicians. High performance scores may result from a small pool of high-performing physicians who report such measures, thereby potentially underestimating the extent of variation of the measure action across physicians. Based on the large discrepancy in the reporting rate for this measure when comparing CMS claims and registry data submissions to the number of eligible reporting instances (based on billed CPT codes for exams relevant to the measure denominator), this measure may have a larger performance gap than the CMS reporting rate shows.

**Table 10. Reporting rate for individuals that submitted to CMS vs. all reporters**

|  |  |  |
| --- | --- | --- |
|  | **Reporting rate across all individuals who reported the measure to CMS** | **Reporting Rate for all eligible reporting instances** |
| 2015 | 88% | 35% |
| 2016 | 90% | 48% |
| 2017 | 86% | 55% |
| 2018 | 91% | 84% |

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

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

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

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

# 2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

# 2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of

missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

With the use of claims and registry as the data sources for this measure, CMS Medicare and Medicaid administrative data is valid and reliable since it determines eligibility for enrollment and payment of services. Registry data submissions may have some missing data, as registry users are not required to submit *all* data to CMS. Registry users are required to submit 70% of their data. However, the volume of patients (6,462,722) used in this data set greatly minimizes the risk of bias.

**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*)

Missing data related to registry data providers not submitting information on patients was previously noted. However, the number of patients that were eligible (7,267,917) compared to the amount submitted and used for this analysis (6,462,722) likely would not have made a significant difference in the testing results. It does, however, make a difference in the performance gap for this measure. The performance may be affected by the lack of responses.

**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*)

The performance results are from a significantly large data set of over 6,000,000 patients. The loss of about 800,000 eligible patients would be unlikely to create a bias or a significant difference in the results. Yearly, CMS raises the volume of data required for submission in the MIPS program. This will assist with minimizing bias even more in the future.