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

**Measure Number** (*if previously endorsed*)**:** 0661

**Measure Title**: Head CT or MRI Scan Results for Acute Ischemic Stroke or Hemorrhagic Stroke Patients who Received Head CT or MRI Scan within 45 Minutes of ED Arrival

**Date of Submission**: 06/30/2015 (Spring 2016 Submission) | 08/01/2019 (Fall 2019 Submission)

**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 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**](#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*).

1. Datasets used to define the initial patient population:

* The initial patient population is identified using chart-abstracted data for a sample of ED encounters with at least one of the following Current Procedural Terminology (CPT) codes for evaluation and management (E/M): 99281, 99282, 99287, 99284, 99285, or 99291. The initial patient population includes cases for patients 18 years and older, as of the date of the encounter, with a principle diagnosis associated with an acute ischemic or hemorrhagic stroke, identified by using any of the following International Classification of Diseases version 9 (ICD-9) codes: 430, 431, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91, or 436.

1. Datasets used to define the denominator:

* The denominator is identified using chart-abstracted data for a sample of cases for patients included in the initial patient population.

1. Datasets used to identify denominator exclusions:

* Denominator exclusions are identified using chart-abstracted data of cases for patients included in the denominator. Denominator exclusions capture cases for patients where any of the following conditions are met:
  + *Discharge Code* is equal or equivalent to “Expired,” “Left Against Medical Advice/AMA” or “not documented or unable to determine (UTD)”
  + *Head CT or MRI Scan Order* is equal to missing or “No”
  + *Last Known Well* is equal to “No”
  + *Time Last Known Well* is greater than 120 minutes

1. Datasets used to capture the numerator:

* The numerator is identified using chart-abstracted data of cases for patients included in the denominator. The numerator includes cases for patients where either of the following conditions are met:
  + *ED Arrival Time* to *Head CT Scan Interpretation Time* is within 45 minutes
  + *ED Arrival Time* to *MRI Scan Interpretation Time* is within 45 minutes

1. Datasets used to identify numerator exceptions:

* Numerator exceptions are identified by using chart-abstracted data of cases for patients included in the denominator. Numerator exceptions include cases of patients for whom any of the following conditions are met:
  + *Date Last Known Well* is equal to “UTD” –*Time Last Known Well* is equal to “UTD,”
  + *Arrival Time* is equal to “UTD”
  + *Head CT or MRI Scan Interpretation Date* is equal to “UTD”
  + *Head CT or MRI Scan Interpretation Time* is equal to “UTD”

1. Datasets used to define the initial patient population:

* The initial patient population is identified using chart-abstracted data for a sample of ED encounters with at least one of the following Current Procedural Terminology (CPT) codes for evaluation and management (E/M):
  + 99281
  + 99282
  + 99283
  + 99284
  + 99285
  + 99291
* The initial patient population includes cases for patients 18 years and older, as of the date of the encounter, with a principle diagnosis for acute ischemic or hemorrhagic stroke, identified by using any of the following International Classification of Diseases, version 10 (ICD-10) codes
  + I6000
  + I6001
  + I6002
  + I6010
  + I6011
  + I6012
  + I602
  + I6030
  + I6031
  + I6032
  + I604
  + I6050
  + I6051
  + I6052
  + I606
  + I607
  + I608
  + I609
  + I610
  + I611
  + I612
  + I613
  + I614
  + I615
  + I616
  + I618
  + I619
  + I629
  + I6300
  + I63011
  + I63012
  + I63013
  + I63019
  + I6302
  + I63031
  + I63032
  + I63033
  + I63039
  + I6309
  + I6310
  + I63111
  + I63112
  + I63113
  + I63119
  + I6312
  + I63131
  + I63132
  + I63133
  + I63139
  + I6319
  + I6320
  + I63211
  + I63212
  + I63213
  + I63219
  + I6322
  + I63231
  + I63232
  + I63233
  + I63239
  + I6329
  + I6330
  + I63311
  + I63312
  + I63313
  + I63319
  + I63321
  + I63322
  + I63323
  + I63329
  + I63331
  + I63332
  + I63333
  + I63339
  + I63341
  + I63342
  + I63343
  + I63349
  + I6339
  + I6340
  + I63411
  + I63412
  + I63413
  + I63419
  + I63421
  + I63422
  + I63423
  + I63429
  + I63431
  + I63432
  + I63433
  + I63439
  + I63441
  + I63442
  + I63443
  + I63449
  + I6349
  + I6350
  + I63511
  + I63512
  + I63513
  + I63519
  + I63521
  + I63522
  + I63523
  + I63529
  + I63531
  + I63532
  + I63533
  + I63539
  + I63541
  + I63542
  + I63543
  + I63549
  + I6359
  + I636
  + I6381
  + I6389
  + I639

1. Datasets used to define the denominator:

* The denominator is identified using chart-abstracted data for a sample of cases for patients included in the initial patient population, who arrived to the emergency department (ED) within 2 hours of the *Time Last Known Well* and had an order for a head CT or MRI scan.

1. Datasets used to identify denominator exclusions:

* Denominator exclusions are identified using chart-abstracted data of cases for patients included in the denominator. Denominator exclusions capture cases for patients where any of the following conditions are met:
  + *Discharge Code* is equal or equivalent to “Expired,” “Left Against Medical Advice/AMA” or “not documented or unable to determine (UTD)”
  + *Head CT or MRI Scan Order* is equal to missing or “No”
  + *Last Known Well* is equal to “No”
  + Difference between *Arrival Time* and *Time Last Known Well* is greater than 120 minutes

1. Datasets used to capture the numerator:

* The numerator is identified using chart-abstracted data of cases for patients included in the denominator. The numerator includes cases for patients where either of the following conditions are met:
  + *ED Arrival Time* to *Head CT Scan Interpretation Time* is within 45 minutes
  + *ED Arrival Time* to *MRI Scan Interpretation Time* is within 45 minutes

1. Datasets used to identify numerator exceptions:

* Numerator exceptions are identified by using chart-abstracted data of cases for patients included in the denominator. Numerator exceptions include cases of patients for whom any of the following conditions are met:
  + *Date Last Known Well* is equal to “UTD” –*Time Last Known Well* is equal to “UTD,”
  + *Arrival Time* is equal to “UTD”
  + *Head CT or MRI Scan Interpretation Date* is equal to “UTD”
  + *Head CT or MRI Scan Interpretation Time* is equal to “UTD”

**1.3. What are the dates of the data used in testing**? January 1, 2014—December 31, 2014 | July 1, 2016–December 31, 2018

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

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

**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 number of measured entities (hospital emergency departments) varies by testing type; see Section **1.7** for details.

The number of measured entities (hospital emergency departments) varies by testing type; see Section **1.7** for details.

**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 number of patients varies by testing type; see Section **1.7** for details.

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**.

**Reliability Testing**

Data Source: Hospital Compare downloadable file [maintained by the Centers for Medicare & Medicaid Services (CMS)]

Dates: *Denominator:* January 1, 2014–December 31, 2014; *Numerator:* January 1, 2014–December 31, 2014; *Exclusions:* January 1, 2014–December 31, 2014; *Exceptions:* January 1, 2014–December 31, 2014

Number of Facilities: 958

Denominator Cases: 17,162

Numerator Cases: 11,741

Level of Analysis: Facility

Patient Characteristics: Not applicable

**Validity Testing—*Empirical Validity***

Data Source: Validation mismatches: Hospital Outpatient Quality Reporting (Hospital OQR) Clinical Data Warehouse (CDW)

Dates: April 1, 2014-March 31, 2015

Sampled Population: 774

Level of Analysis: Data element

Patient Characteristics: Not applicable

**Validity Testing—*Face Validity***

Data Source: Structured qualitative survey completed by the stroke and acute myocardial infarction expert work group (EWG) members

Date Collected: October - November 2015

Number of Responses: 5

Respondent Characteristics: Respondents were asked to self-identify as one or more of the following categories: clinician (4); healthcare administration (1); management (1), other—payer consultant (1), other—Institute Director, Association (1)

**Exclusions Analysis**

Data Source: Denominator: CDW; Numerator: CDW; Exclusions: CDW

Dates: *Denominator:* January 1, 2014–December 31, 2014; *Numerator:* January 1, 2014–December 31, 2014; *Exclusions:* January 1, 2014–December 31, 2014; *Exceptions:* January 1, 2014–December 31, 2014

Number of Facilities: 3,614

Sampled Population: 105,898

Denominator Cases: 28,236

Numerator Cases: 18,480

Level of Analysis: Case

Denominator Patient Characteristics: Gender (% Male): 49.0; Mean Age (Years): 66.8 (St. Dev.: 15.3); Race (% Minority): 19.3

**Risk Adjustment/Stratification**

N/A—No risk adjustment or stratification was performed.

**Identification of Statistically Significant & Meaningful Differences in Performance**

Data Source: Hospital Compare downloadable file

Dates: *Denominator:* January 1, 2014–December 31, 2014; *Numerator:* January 1, 2014–December 31, 2014; *Exclusions:* January 1, 2014–December 31, 2014; *Exceptions:* January 1, 2014–December 31, 2014

Number of Facilities: 958

Denominator Cases: 17,162

Numerator Cases: 11,741

Level of Analysis: Facility

Patient Characteristics: Not applicable

**Comparability of Performance Scores when more than one Set of Specifications**

N/A—This measure only uses one set of specifications.

**Missing Data Analysis and Minimizing Bias**

Data Source: Denominator: CDW; Numerator: CDW; Exclusions: CDW

Dates: *Denominator:* January 1, 2014–December 31, 2014; *Numerator:* January 1, 2014–December 31, 2014; *Exclusions:* January 1, 2014–December 31, 2014; *Exceptions:* January 1, 2014–December 31, 2014

Number of Facilities: 3,614

Sampled Population: 105,898

Denominator Cases: 28,236

Numerator Cases: 18,480

Level of Analysis: Case

Denominator Patient Characteristics: Gender (% Male): 49.0; Mean Age (Years): 66.8 (St. Dev.: 15.3); Race (% Minority): 19.3

**Reliability Testing**

Data Source: Hospital Compare downloadable file [maintained by the Centers for Medicare & Medicaid Services (CMS)]

Dates: *Denominator:* July 1, 2017–June 30, 2018; *Numerator:* July 1, 2017–June 30, 2018; *Exclusions:* July 1, 2017–June 30, 2018; *Exceptions:* July 1, 2017–June 30, 2018

Number of Facilities: 1,550

Denominator Cases: 31,939

Numerator Cases: 23,953

Level of Analysis: Facility

Patient Characteristics: Not applicable

**Validity Testing—*Empirical Validity***

Data Source: Clinical Data Abstraction Center (CDAC), Clinical Data Warehouse (CDW)

Dates: July 1, 2016–September 30, 2018

Sampled Population: 2,622

Level of Analysis: Data element

Patient Characteristics: Not applicable

**Exclusions Analysis**

Data Source: Denominator: CDW; Numerator: CDW; Exclusions: CDW

Dates: *Denominator:* July 1, 2016—December 31, 2018; *Numerator:* July 1, 2016—December 31, 2018; *Exclusions:* July 1, 2016—December 31, 2018; *Exceptions:* July 1, 2016—December 31, 2018

Number of Facilities: 3,140

Sampled Population: 271,756

Denominator Cases: 271,756

Numerator Cases: 80,749

Level of Analysis: Case

Denominator Patient Characteristics: Gender (% Male): 49.7; Mean Age (Years): 67.1 (St. Dev.: 15.3); Race (% Minority): 16.4

**Risk Adjustment/Stratification**

N/A—No risk adjustment or stratification was performed.

**Identification of Statistically Significant & Meaningful Differences in Performance**

Data Source: Hospital Compare downloadable file

Dates: *Denominator:* July 1, 2017–June 30, 2018; *Numerator:* July 1, 2017–June 30, 2018; *Exclusions:* July 1, 2017–June 30, 2018; *Exceptions:* July 1, 2017–June 30, 2018

Number of Facilities: 1,550

Denominator Cases: 31,939

Numerator Cases: 23,953

Level of Analysis: Facility

Patient Characteristics: Not applicable

**Comparability of Performance Scores when more than one Set of Specifications**

N/A—This measure only uses one set of specifications.

**Missing Data Analysis and Minimizing Bias**

Data Source: Hospital Compare downloadable file

Dates: *Denominator:* July 1, 2017–June 30, 2018; *Numerator:* July 1, 2017–June 30, 2018; *Exclusions:* July 1, 2017–June 30, 2018; *Exceptions:* July 1, 2017–June 30, 2018

Number of Facilities: 1,550

Denominator Cases: 31,939

Numerator Cases: 23,953

Level of Analysis: Facility

Patient Characteristics: 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.

We assessed patient-level SDS factors as part of the regression model reported in Section **1b.4**, which provides an overview of disparities in care for patient sub-populations. We based this analysis on SDS variables included in the CDW data:

* Age
* Sex
* Race
* Ethnicity

While an analysis of SDS factors is important in understanding differences in care for patient sub-populations, this measure is a process measure that is neither risk-adjusted nor risk-stratified. We determined that risk adjustment and risk stratification were not appropriate based on the current evidence base and the measure construct. Additional information on this determination is provided in Section **2b4.2**.

We assessed patient-level sociodemographic status (SDS) factors as part of the regression model reported in Section **2b3.2**, which provides an overview of disparities in care for patient sub-populations. We based this analysis on SDS variables included in the CDW data:

* Age
* Sex
* Race
* Ethnicity

While an analysis of SDS factors is important in understanding differences in care for patient sub-populations, this measure is a process measure that is neither risk-adjusted nor risk-stratified. We determined that risk adjustment and risk stratification were not appropriate based on the current evidence base and the measure construct. Additional information on this determination is provided in Section **2b3.2.**

**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 was calculated in accordance with the methods discussed in *The Reliability of Provider Profiling: A Tutorial* (2009). This approach calculates the ability of the measure to distinguish between the performances of different facilities. Specifically, the testing calculated the signal-to-noise ratio for each facility meeting the minimum case count, established by the measure calculation contractor, during the 2014 data collection period, with higher scores indicating greater reliability. The reliability score is estimated using a beta-binomial model, which is appropriate for the reliability testing of pass/fail measures. The reliability score for each facility is a function of the facility’s sample size and score on the measure, and the variance across facilities.

REFERENCES:

Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation. 2009. Retrieved from <http://www.rand.org/pubs/technical_reports/TR653>.

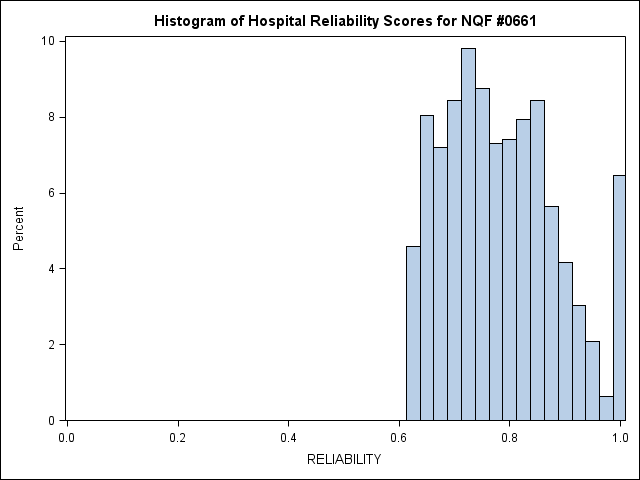
Reliability was calculated in accordance with the methods discussed in *The Reliability of Provider Profiling: A Tutorial* (2009). This approach calculates the ability of the measure to distinguish between the performances of different facilities. Specifically, the testing calculated the signal-to-noise ratio for each facility meeting the minimum case count, established by the measure calculation contractor, during the July 1, 2017–June 30, 2018 data collection period, with higher scores indicating greater reliability. The reliability score is estimated using a beta-binomial model, which is appropriate for the reliability testing of pass/fail measures. The reliability score for each facility is a function of the facility’s sample size and score on the measure, and the variance across facilities.

REFERENCES:

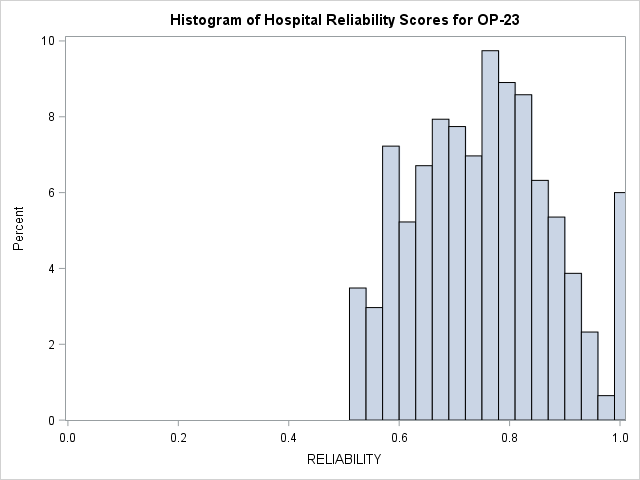
Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation. 2009. Retrieved from <http://www.rand.org/pubs/technical_reports/TR653>.

**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*)

*Figure 1* (below) is a histogram of the distribution of the reliability scores for the facilities meeting the minimum case count requirements during the 2014 data collection period. Reliability scores ranged from 0.62 to 1.00, with a median reliability score of 0.77. Higher scores denote greater reliability.



*Figure 1* (below) is a histogram of the distribution of the reliability scores for the facilities meeting the minimum case count requirements during the July 1, 2017 – June 30, 2018 data collection period. Reliability scores ranged from 0.52 to 1.00, with a median reliability score of 0.76. Higher scores denote greater reliability.



**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Calculated using a beta-binomial model, a median reliability score of 0.77 is indicative of strong measure reliability. The results of this test indicate that the measure is able to identify true differences in performance between individual facilities.

Calculated using a beta-binomial model, a median reliability score of 0.76 is indicative of strong measure reliability. The results of this test indicate that the measure is able to identify true differences in performance between individual facilities.

**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)*

We assessed the validity of the measure using both qualitative and quantitative analyses. Empirical validity of the data elements was assessed by calculating a rate of agreement between facility abstraction and auditor (CDAC) abstraction for each of the data elements used to calculate the measure. The empirical analysis used data element values for 774 cases abstracted by CDAC, which were previously abstracted by facilities. The data was collected from April 1, 2014 to March 31, 2015.

Face validity of the measure score was systematically assessed through survey of the EWG. Five EWG members participated in the data collection. Respondent perspectives include clinicians, management, and healthcare administration. Prior to responding to questions related to measure-score face validity, EWG members were provided detailed measure specifications.

The following questions and statements related to measure-score face validity were posed to the EWG:

1. Patients who have a head CT scan or MRI ordered and interpreted within 45 minutes of ED arrival can be accurately captured using chart-abstracted data.
2. The measure successfully assesses the timely interpretation of head imaging for acute ischemic and hemorrhagic stroke patients.

Responses to questions *1* and *2* in the measure-score face-validity section were collected using a five-point Likert scale: strongly agree, agree, undecided, disagree, strongly disagree, and do not know.

Empirical validity of critical data elements was assessed by examining kappa statistics (for categorical variables and the constructed outcomes of the numerator and denominator) and Pearson’s correlation coefficient (for non-categorical variables) between facility abstraction and auditor (CDAC) abstraction for each of the data elements used to calculate the measure. This analysis used data elements for 2,622 cases abstracted by CDAC, which were previously abstracted by facilities; these data were collected from July 1, 2016 to December 31, 2018.

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

Results of empirical validity testing indicate moderate to strong levels of agreement between the facility’s abstraction of data elements versus CDAC’s abstraction of data elements for the same sample of cases. The rate of agreement, by data element, ranged from 52.7% to 98.4%.

Results of the face-validity assessment indicate that a diverse group of stakeholders support the validity of the measure. Results for each of the questions provided above follow.

1. *Patients who have a head CT scan or MRI ordered and interpreted within 45 minutes of ED arrival can be accurately captured using chart-abstracted data.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 60.0% | 3 |
| *Agree* | 40.0% | 2 |
| *Undecided* | 0.0% | 0 |
| *Disagree* | 0.0% | 0 |
| *Strongly Disagree* | 0.0% | 0 |
| *Do Not Know or Not Applicable* | 0.0% | 0 |

1. *The measure successfully assesses the timely interpretation of head imaging for acute ischemic and hemorrhagic stroke patients.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 40.0% | 2 |
| *Agree* | 60.0% | 3 |
| *Undecided* | 0.0% | 0 |
| *Disagree* | 0.0% | 0 |
| *Strongly Disagree* | 0.0% | 0 |
| *Do Not Know or Not Applicable* | 0.0% | 0 |

Results of the data element validity analysis are presented and organized by variable type in the tables below.

*Table 1.* Data Element Validity for Categorical Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable (Categorical)** | **Kappa Value** | **p Value** | **n** |
| *Discharge* | 0.9318 | 0.0000 | 2,617 |
| *Head CT/MRI Scan* | 0.8625 | 0.0000 | 2,471 |
| *Last Known Well* | 0.7676 | 0.0000 | 2,394 |

*Table 2.* Data Element Validity for Non-Categorical Variables

| **Variable (Non-Categorical)** | **Pearson's Correlation Coefficient** | **p Value** | **n** |
| --- | --- | --- | --- |
| *Date Last Known Well (LKW)* | 1.0000 | p < .001 | 1,391 |
| *Date of CT/MRI Scan* | 1.0000 | p < .001 | 668 |
| *Time of Arrival* | 0.5122 | p < .001 | 95 |
| *Time LKW* | 0.9288 | p < .001 | 1,386 |
| *Time of CT/MRI Scan* | 0.9882 | p < .001 | 659 |

*Table 3.* Validity of Constructed Outcomes

| **Variable (Categorical)** | **Kappa Value** | **p Value** | **n** |
| --- | --- | --- | --- |
| *Numerator* | 0.8450 | 0.0000 | 2,622 |
| *Denominator* | 0.8464 | 0.0000 | 2,622 |

**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?*)

Results of the quantitative and qualitative analysis are positive and support the validity of the measure and its calculation. The rate of agreement between facility and CDAC abstraction ranged from moderate to strong across the data elements used to calculate OP-23. The rate of agreement was strong for dichotomous variables, as well as those based on administrative data. Agreement was moderate for clinical variables related to time.

The EWG, composed of five stakeholders representing healthcare administration, management, payer consultants, associations, and clinicians with expertise in cardiology, neuro-radiology, emergency medicine, and emergency nursing, provided feedback on the face validity of NQF #0661 through an online survey. All members agreed or strongly agreed that patients who have a head CT scan or MRI ordered and interpreted within 45 minutes of ED arrival can be accurately captured using chart-abstracted data. Similarly, they agreed or strongly agreed that NQF #0661 successfully assesses the timely interpretation of head imaging for acute ischemic and hemorrhagic stroke patients. The respondents generally support the face validity of NQF #0661.

The agreement between facility and CDAC-abstracted data elements ranged from moderate to strong across the data elements used to calculate OP-23. Kappa values ranged from 0.77–0.93 for categorical data elements; Pearson’s correlation coefficients for non-categorical variables ranged from 0.51-0.92; and Kappa values for the constructed variables of the numerator and denominator were each 0.85. These results support the validity of the measure and its calculation.

**2b2. EXCLUSIONS ANALYSIS**

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

**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*)

We tested measure exclusions and numerator exceptions to determine the prevalence of each exclusion and exception, by facility, and at an aggregate level. The analysis tested measure exclusions and numerator exceptions during the 2014 data collection period. Measure exclusions include all cases meeting one or more criteria listed in Section **1.2c**, above. Numerator exceptions include cases meeting one or more criteria listed in Section **1.2d**, above. To supplement the empirical results, we systematically assessed the face validity of current exclusions through survey of the EWG based on responses from five EWG members.

The face validity of exclusions was assessed, using the following questions and statements:

1. To be included in the measure population, each patient must receive care in the ED for a stroke. These patients are identified based on ICD-9 principal diagnosis codes and E&M codes. From this initial patient population, certain patients are excluded from NQF #0661 based on the situations listed in the table below.[[1]](#footnote-1) Please evaluate the appropriateness of each of the CURRENT exclusion criteria.
2. For NQF #0661, do you foresee any challenges in capturing any of the exclusions using chart-abstracted data?

Responses to question *1* were collected using keep/remove response options; responses to question *2* were collected using yes/no response options.

We tested measure exclusions and numerator exceptions to determine the prevalence of each exclusion and exception, by facility, and overall. The analysis tested measure exclusions and numerator exceptions during the July 1, 2016—December 31, 2018 data collection period. Measure exclusions include all cases meeting one or more criteria listed in Section **1.2c**, above. Numerator exceptions include cases meeting one or more criteria listed in Section **1.2e**, above.

**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*)

We examined overall frequencies and proportions of cases excluded for each exclusion/exception criterion, among all sampled cases, for 3,614 facilities submitting eligible cases in 2014. The sampled population included 105,898 cases where a patient (age 18 years or older) presented with an ischemic or hemorrhagic stroke to an ED.

Overall Occurrence and Distribution across Facilities for Measure Exclusions and Exceptions

| **Data Element** | **Denominator Exclusion or Numerator Exception?** | | **Overall Occurrence** | | **Distribution Across Facilities (%)** | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Denominator Exclusion** | **Numerator Exception** | **N** | **%** | **25th** | **50th** | **75th** |
| *Discharge Code* | X |  | 4,627 | 4.4 | 0.0 | 2.5 | 6.7 |
| *Head CT Scan or MRI Order* | X |  | 3,537 | 3.3 | 0.0 | 1.6 | 5.3 |
| *Last Known Well* | X |  | 44,328 | 41.9 | 28.6 | 41.4 | 52.6 |
| *Date Last Known Well* |  | X | 100 | 0.1 | 0.0 | 0.0 | 0.0 |
| *Time Last Known Well* |  | X | 1,506 | 1.4 | 0.0 | 0.0 | 0.0 |
| *Arrival Time* |  | X | 5 | 0.0 | 0.0 | 0.0 | 0.0 |
| *Last Known Well Minutes* | X |  | 25,170 | 23.8 | 14.3 | 22.2 | 30.4 |
| *Head CT Scan or MRI Interpretation Date* |  | X | 87 | 0.1 | 0.0 | 0.0 | 0.0 |
| *Head CT Scan or MRI Interpretation Time* |  | X | 373 | 0.4 | 0.0 | 0.0 | 0.0 |
|  | | | | | | | |
| *Total Denominator Exclusions* | 4 exclusions | - | 77,662 | 73.40 | - | - | - |
| *Total Numerator Exceptions* | - | 5 exceptions | 2,071 | 2.00 | - | - | - |
|  | | | | | | | |
| *Total Removed from the Denominator or Numerator* | 9 exceptions and exclusions | | 79,733 | 75.40 | - | - | - |

As indicated in the table above, 73.4% of the initial patient population is excluded from the denominator. This may be partially explained by clinical factors, such as stroke patients presenting to the ED after two hours of symptom onset, or by limitations of chart abstracted documentation, which might have a large volume of missing and/or abnormal cases. However, the use of minimum case counts ensures that we report performance scores for facilities that have an adequate number of cases after the application of these exclusions.

We examined overall frequencies and proportions of cases excluded for each exclusion/exception criterion, among all sampled cases, for 3,140 facilities submitting eligible cases from July 1, 2016—December 31, 2018. The sampled population included 271,756 cases where a patient (ages 18 years or older) presented with an ischemic or hemorrhagic stroke to an ED.

*Table 4.* Overall Occurrence and Distribution across Facilities for Measure Exclusions and Exceptions

| **Data Element** | **Denominator Exclusion or Numerator Exception?** | | **Overall Occurrence** | | **Distribution Across Facilities (%)** | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Denominator Exclusion** | **Numerator Exception** | **N** | **%** | **25th** | **50th** | **75th** |
| *Discharge Code* | X |  | 10,702 | 3.94 | 2.0 | 4.0 | 7.0 |
| *Head CT Scan or MRI Order* | X |  | 7,635 | 2.92 | 1.0 | 2.0 | 5.0 |
| *Last Known Well* | X |  | 101,127 | 37.21 | 27.0 | 44.0 | 69.0 |
| *Date Last Known Well* |  | X | 0 | 0.00 | 0.0 | 0.0 | 0.0 |
| *Time Last Known Well* |  | X | 192 | 0.00 | 0.0 | 0.0 | 0.0 |
| *Arrival Time* |  | X | 0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *Last Known Well Minutes* | X |  | 77,757 | 39.38 | 22.0 | 34.0 | 52.0 |
| *Head CT Scan or MRI Interpretation Date* |  | X | 0 | 0.00 | 0.0 | 0.0 | 0.0 |
| *Head CT Scan or MRI Interpretation Time* |  | X | 54 | 0.00 | 0.0 | 0.0 | 0.0 |
|  |  |  |  |  |  |  |  |
| *Total Denominator Exclusions* | 4 exclusions | - | 197,221 | 72.66 | - | - | - |
| *Total Numerator Exceptions* | - | 5 exceptions | 246 | 0.00 | - | - | - |
|  |  |  |  |  |  |  |  |
| *Total Removed from the Denominator or Numerator* | 9 exceptions and exclusions | | 197,467 | 72.66 | - | - | - |

As indicated in the table above, 72.7% of the initial patient population is excluded from the denominator. However, the use of minimum case counts ensures that we report performance scores for facilities that have an adequate number of cases after the application of these exclusions.

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

As seen in the table reported in Section **2b3.2** above, the frequency of exclusions/exceptions varied substantially across facilities. Measure exclusion and numerator exception criteria are in alignment with clinical guidelines and also ensure that all cases included in the measure have sufficient denominator and numerator information to calculate the performance score. After identification of cases for patients 18 years and older with a principal diagnosis associated with acute ischemic or hemorrhagic stroke, measure exclusion and numerator exception criteria are applied:

* *Discharge Code* is a measure exclusion criterion. Cases for patients where *Discharge Code* equals “Expired”, “Left Against Medical Advice/AMA”, or “UTD” are excluded from the measure denominator. Overall, 4.4% of cases for patients included in the initial patient population are excluded from the denominator based on *Discharge Code*. There is notable variability in the proportion of cases excluded based on *Discharge Code* values across facilities, with an inter-quartile range of 0.0% to 6.7%.
* *Head CT Scan or MRI Order* is a measure exclusion criterion. Cases for patients where *Head CT Scan or MRI Order* equals “No” are excluded from the denominator. This criterion is based off the fact that the numerator is dependent upon a head CT or MRI scan being performed. Overall, 3.3% of cases included in the initial patient population, were excluded from the measure denominator based on *Head CT Scan or MRI Order*. There is notable variability in the proportion of excluded *Head CT Scan or MRI Order* values across facilities, with an interquartile range from 0.0% to 5.3%.
* *Last Known Well* is a measure exclusion criterion. It is a binary variable that indicates if there are values for both *Date Last Known Well* and *Time Last Known Well.* Cases for patients where *Last Known Well* is equal to “No” are excluded from the measure denominator. Overall, 41.9% of cases for patients included in the initial patient population had a *Last Known Well* value equal to “No.” There is large variability in the proportion of excluded *Last Known Well* values across facilities, with an interquartile range from 28.6% to 52.6%.
* *Date Last Known Well* is a numerator exception criterion. If *Date Last Known Well* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, 0.1% of patients included in the denominator have a “UTD” value for *Date Last Known Well*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Time Last Known Well* is a numerator exception criterion. If *Time Last Known Well* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, 1.4% of patients included in the denominator have a “UTD” value for *Time Last Known Well*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Arrival Time* is a numerator exception criterion. If *Arrival Time* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, less than 0.1% of patients included in the denominator have a “UTD” value for *Arrival Time*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Last Known Well Minutes* is a measure exclusion criterion. It is a measurement value calculated from *Arrival Time* and *Time Last Known Well.* If *Last Known Well Minutes* is greater than 120 minutes, the case is excluded from the measure. This criterion is based off of clinical guidelines for the most appropriate time window to identify and treat acute stroke. Overall, 23.8% of patients eligible for the measure have a value greater than 120 minutes for *Time Last Known Well Minutes; although this value may appear high, multiple studies have found* *that fewer than one-third of stroke patients arrive at the ED within two hours of symptom onset (Mozaffarian et al. 2015; Pittenger et al. 2014)*. There is variability in the proportion of excluded values for *Last Known Well Minutes* across facilities, with an interquartile range of 14.3% to 30.4%.
* *Head CT Scan or MRI Interpretation Date* is a numerator exception criterion. If *Head CT Scan or MRI Interpretation Date* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, 0.1% of patients included in the denominator have a “UTD” value for *Head CT Scan or MRI Interpretation Date*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Head CT Scan or MRI Interpretation Time* is a numerator exception criterion. If *Head CT Scan or MRI Interpretation Time* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, 0.4% of patients included in the denominator have a “UTD” value for *Head CT Scan or MRI Interpretation Date*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.

Results of the survey of the EWG also support the face validity of the exclusions and exceptions for NQF #0661, and indicate that these exclusions are consistent with prevailing gold standards of care or are necessary to support measure calculation.

As seen in table **2b2.2** above, a large number of patients (72.7%) are excluded from the measure calculation; the removal of cases where an abstractor submits a value of “UTD” for *Last Known Well*  is necessary to align with clinical guidelines and enable measure calculation. Missing data on this criterion makes it impossible to determine the timeliness of interpretation of a head CT or MRI. The frequency of exclusions/exceptions varied substantially across facilities. Measure exclusion and numerator exception criteria are in alignment with clinical guidelines and also ensure that all cases included in the measure have sufficient denominator and numerator information to calculate the performance score. After identification of cases for patients 18 years and older with a principal diagnosis associated with acute ischemic or hemorrhagic stroke, measure exclusion and numerator exception criteria are applied:

* *Discharge Code* is a measure exclusion criterion. Cases for patients where *Discharge Code* equals “Expired”, “Left Against Medical Advice/AMA”, or “UTD” are excluded from the measure denominator. Overall, 3.94% of cases for patients included in the initial patient population are excluded from the denominator based on *Discharge Code*. There is notable variability in the proportion of cases excluded based on *Discharge Code* values across facilities, with an interquartile range of 2.0% to 7.0%.
* *Head CT Scan or MRI Order* is a measure exclusion criterion. Cases for patients where *Head CT Scan or MRI Order* equals “No” are excluded from the denominator. This criterion is based off the fact that the numerator is dependent upon a head CT or MRI scan being performed. Overall, 3.3% of cases included in the initial patient population, were excluded from the measure denominator based on *Head CT Scan or MRI Order*. There is notable variability in the proportion of excluded *Head CT Scan or MRI Order* values across facilities, with an interquartile range from 1.0 to 5.0%.
* *Last Known Well* is a measure exclusion criterion. It is a binary variable that indicates if there are values for both *Date Last Known Well* and *Time Last Known Well.* Patients where *Last Known Well* is equal to “No” are excluded from the measure denominator. Overall, 37.21% of cases for patients included in the initial patient population had a *Last Known Well* value equal to “No.” There is large variability in the proportion of excluded *Last Known Well* values across facilities, with an interquartile range from 27.0% to 69.0%.
* *Date Last Known Well* is a numerator exception criterion. If *Date Last Known Well* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. No cases of patients included in the denominator had “UTD” value for *Date Last Known Well*. The exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Time Last Known Well* is a numerator exception criterion. If *Time Last Known Well* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, <0.1% of patients included in the denominator have a “UTD” value for *Time Last Known Well*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Arrival Time* is a numerator exception criterion. If *Arrival Time* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. No patients included in the denominator have a “UTD” value for *Arrival Time*. The exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Last Known Well Minutes* is a measure exclusion criterion. It is a measurement value calculated from *Arrival Time* and *Time Last Known Well.* If *Last Known Well Minutes* is greater than 120 minutes, the case is excluded from the measure. This criterion is based off of clinical guidelines for the most appropriate time window to identify and treat acute stroke. Overall, 28.61% of patients eligible for the measure have a value greater than 120 minutes for *Time Last Known Well Minutes; although this value may appear high, multiple studies have found* *that fewer than one-third of stroke patients arrive at the ED within two hours of symptom onset (Mozaffarian et al. 2015; Pittenger et al. 2014)*. There is variability in the proportion of excluded values for *Last Known Well Minutes* across facilities, with an interquartile range of 22.0% to 52.0%.
* *Head CT Scan or MRI Interpretation Date* is a numerator exception criterion. If *Head CT Scan or MRI Interpretation Date* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. No patients included in the denominator have a “UTD” value for *Head CT Scan or MRI Interpretation Date*. The exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.
* *Head CT Scan or MRI Interpretation Time* is a numerator exception criterion. If *Head CT Scan or MRI Interpretation Time* is equal to “UTD,” the case is not included in the measure numerator but remains in the measure denominator. Overall, <0.1% of patients included in the denominator have a “UTD” value for *Head CT Scan or MRI Interpretation Date*. The exception remains important as a “UTD” value for this data element makes it impossible to determine whether the patient received timely interpretation of their head CT or MRI.

**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.

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**.

This measure is a process measure for which we provide no risk adjustment or stratification. We determined risk adjustment and stratification were not appropriate based on the measure evidence base and the measure construct. As a process-of-care measure, the decision to order and interpret a head CT or MRI scan within 45 minutes should not be influenced by SDS factors; rather, adjustment would potentially mask such important inequities in care delivery. Variation across patient populations is reflective of differences in the quality of care provided to the disparate patient population included in the measure’s denominator.

During the measure maintenance process, we perform an annual review of the literature, to identify articles and clinical practice guidelines published in the last 12 months, which includes a scan for potential patient subpopulations for which there are differences in the clinical decision to perform a head CT or MRI scan; this most recent review identified no clear evidence of an empirical relationship between SDS and facility-level measure performance.

In addition to the evidence gathered from the literature, stakeholder feedback obtained during the three years of implementation and public reporting has not identified concerns related to SDS factors and need for risk adjustment. This supports the conceptual model upon which the measure is based.

This measure is a process measure for which we provide no risk adjustment or stratification. We determined risk adjustment and stratification were not appropriate based on the measure evidence base and the measure construct. As a process-of-care measure, the decision to order and interpret a head CT or MRI scan within 45 minutes should not be influenced by SDS factors; rather, adjustment would potentially mask such important inequities in care delivery. Variation across patient populations is reflective of differences in the quality of care provided to the disparate patient population included in the measure’s denominator.

Results of logistic regression of inclusion in the measure’s numerator (i.e., having timely order and interpretation of head CT or MRI) on patient characteristics are provided in *Table 5*, below. Odds of inclusion in the measure’s numerator are significantly lower for beneficiaries in younger age ranges as compared to the referent group (60 to 69); whereas, odds of inclusion are marginally higher (for 70 to 79) or significantly higher (for 80 to 89 and 90 and older) for older age ranges. Compared to males, females were less likely to be included in the numerator. Race was not associated with odds of inclusion in the numerator; however, Hispanic/Latino ethnicity was associated with lower odds of inclusion in the numerator. Compared with the smallest facilities (i.e., those with 50 or fewer beds) beneficiaries at larger facilities up to 500 beds had higher odds of numerator inclusion. Facility location was not associated with odds of inclusion in the numerator. Compared to non-teaching facilities, beneficiaries at major teaching facilities had significantly lower odds of inclusion in the numerator.

*Table 5.* Logistic Regression Results

| *Variable* | *Odds Ratio* | *CI Lower Bound* | *CI Upper Bound* | *P Value* |
| --- | --- | --- | --- | --- |
| *Beneficiary Age* |  |  |  |  |
| 18 to 29 | 0.460 | 0.400 | 0.530 | 0.000 |
| 30 to 39 | 0.612 | 0.555 | 0.674 | 0.000 |
| 40 to 49 | 0.773 | 0.724 | 0.825 | 0.000 |
| 50 to 59 | 0.942 | 0.892 | 0.995 | 0.032 |
| 60 to 69 (Referent) | -- | *--* | *--* | *--* |
| 70 to 79 | 1.051 | 0.999 | 1.106 | 0.057 |
| 80 to 89 | 1.192 | 1.127 | 1.261 | 0.000 |
| 90 and Older | 1.196 | 1.097 | 1.304 | 0.000 |
| *Beneficiary Gender* |  |  |  |  |
| Male (Referent) | *--* | *--* | *--* | *--* |
| Female | 0.874 | 0.845 | 0.905 | 0.000 |
| Unknown | 1.661 | 0.388 | 7.113 | 0.494 |
| *Beneficiary Race* |  |  |  |  |
| Caucasian (Referent) | *--* | *--* | *--* | *--* |
| African-American | 0.942 | 0.878 | 1.010 | 0.095 |
| Asian | 0.985 | 0.844 | 1.150 | 0.850 |
| America Indian or Alaska Native | 0.813 | 0.653 | 1.012 | 0.064 |
| Native Hawaiian or Pacific Islander | 0.826 | 0.590 | 1.156 | 0.265 |
| Unknown Racial or Ethnic Identification | 0.917 | 0.795 | 1.057 | 0.231 |
| *Ethnicity* |  |  |  |  |
| Non-Hispanic/Latino (referent) | -- | -- | -- | -- |
| Hispanic/Latino | 0.805 | 0.707 | 0.915 | 0.001 |
| *Facility Bed Count* |  |  |  |  |
| 0 to 50 Beds (Referent) | -- | -- | -- | -- |
| 51 to 100 Beds | 1.187 | 1.066 | 1.322 | 0.002 |
| 101 to 250 Beds | 1.353 | 1.221 | 1.499 | 0.000 |
| 251 to 500 Beds | 1.348 | 1.195 | 1.521 | 0.000 |
| 500+ Beds | 0.876 | 0.744 | 1.030 | 0.109 |
| Unknown Bed Count | - | - | - | - |
| *Facility Location* |  |  |  |  |
| Rural (Referent) | -- | -- | -- | -- |
| Urban | 0.991 | 0.912 | 1.076 | 0.821 |
| Location Unknown | 1.654 | 1.113 | 2.456 | 0.013 |
| *Facility Teaching Status* |  |  |  |  |
| Non-Teaching (Referent) | -- | -- | -- | -- |
| Teaching | 1.046 | 0.936 | 1.168 | 0.426 |
| Major Teaching | 0.770 | 0.627 | 0.946 | 0.013 |
| Unknown Teaching Status | 0.394 | 0.344 | 0.452 | 0.000 |
| *Constant* | 2.958 | 2.712 | 3.226 | 0.000 |

During the measure maintenance process, we perform an annual review of the literature, to identify articles and clinical practice guidelines published in the last 12 months, which includes a scan for potential patient subpopulations for which there are differences in the clinical decision to perform a head CT or MRI scan; this most recent review identified no clear evidence of an empirical relationship between SDS and facility-level measure performance.

In addition to the evidence gathered from the literature, stakeholder feedback obtained during the three years of implementation and public reporting has not identified concerns related to SDS factors and need for risk adjustment. This supports the conceptual model upon which the measure is based.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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*)

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

*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)

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

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

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

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

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

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

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

Not applicable—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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)*

We tested the statistical significance of the difference between facility performance scores and the mean performance value for facilities meeting public-reporting requirements. For the 2014 data, this included 958 facilities. For each facility, the facility performance score and standard deviation was calculated. This analysis identified 43 facilities as statistical outliers. Additional details of this analysis are provided in Section **2b5.2**.

We tested the statistical significance of the difference between facility performance scores and the mean performance value for facilities meeting public-reporting requirements. For the July 1, 2017–June 30, 2018 data, this included 1550 facilities. For each facility, the facility performance score and standard deviation was calculated. This analysis identified 92 facilities as statistical outliers. Additional details of this analysis are provided in Section **2b4.2**.

**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*)

Of the 958 facilities reporting during the 2014 data collection period, 43 (4.5%) facilities had a performance value that was statistically significantly different from a mean benchmark value. Statistically meaningful difference was defined as when the facility score fell outside of the confidence interval (± 1.96 standard deviations) for the measure mean (benchmark value). Thus, this calculation identifies statistical outliers.

Of the 1,550 facilities reporting during the July 1, 2017–June 30, 2018 data collection period, 92 (5.9%) facilities had a performance value that was statistically significantly different from a mean benchmark value. Statistically meaningful difference was defined as when the facility score fell outside of the confidence interval (± 1.96 standard deviations) for the measure mean (benchmark value). Thus, this calculation identifies statistical outliers. *Table 6*, below, displays the distribution of performance scores for OP-23 across common percentiles. The distribution of performance scores, combined with the identification of outlier facilities, demonstrates the ability of the measure to meaningfully distinguish differences in facility performance.

*Table 6.* Distribution of Performance Scores (among Facilities Meeting Minimum Case Count)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Mean (SD)* | *1%* | *5%* | *10%* | *25%* | *50%* | *75%* | *90%* | *95%* | *99%* |
| *Performance* | 77.11% (14.84) | 28.57 | 48.48 | 57.14 | 69.57 | 80.00 | 87.50 | 92.94 | 95.65 | 100.00 |

**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?*)

Analysis of the 2014 performance data, and the subsequent rate of identification of statistically different performance for 4.5% of measured entities, demonstrates the ability of the measure to identify outlying performance. By reporting a measure mean (benchmark value), this provides an opportunity for outlying facilities to identify underperformance related to delayed interpretation of head CT or MRI and work to implement quality improvement mechanisms to increase the proportion of patients receiving rapid interpretation of head CT or MRIs scans when clinically appropriate.

Analysis of the July 1, 2017–June 30, 2018 performance data, and the subsequent rate of identification of statistically different performance for 4.5% of measured entities, demonstrates the ability of the measure to identify outlying performance. By reporting a measure mean (benchmark value), this provides an opportunity for outlying facilities to identify underperformance related to delayed interpretation of head CT or MRI and work to implement quality improvement mechanisms to increase the proportion of patients receiving rapid interpretation of head CT or MRIs scans when clinically appropriate.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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—No risk adjustment or stratification was performed.

Not applicable—No risk adjustment or stratification was performed.

**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*)

This measure is calculated using chart-abstracted data. To limit the effects of missing data, abstractors cannot submit a value of “missing” for individual data elements. When they submit a value of “missing” the case is rejected from the abstraction tool. While abstractors cannot submit missing data, they may submit a value of “UTD” for select data elements for which missing information may be more likely, for example for *Time Last Known Well* before the onset of stroke symptoms. Cases where a value of “UTD” affects clinical decision making are excluded from the measure. Cases where a value of “UTD” is reflective of poor documentation are included in the denominator but excepted from the numerator. To identify the extent and distribution of cases with a value of “UTD” for a data element, we calculated the frequency of such cases as well as the distribution of cases across eligible facilities. The frequency and distribution of missing data are described in Section **2b3.3.**

This measure is calculated using chart-abstracted data. To limit the effects of missing data, abstractors cannot submit a value of “missing” for individual data elements. When they submit a value of “missing” the case is rejected from the abstraction tool. While abstractors cannot submit missing data, they may submit a value of “UTD” for select data elements for which missing information may be more likely, for example for *Time Last Known Well* before the onset of stroke symptoms. Cases where a value of “UTD” affects clinical decision making are excluded from the measure. Cases where a value of “UTD” is reflective of poor documentation are included in the denominator but excepted from the numerator. To identify the extent and distribution of cases with a value of “UTD” for a data element, we calculated the frequency of such cases as well as the distribution of cases across eligible facilities. The frequency and distribution of missing data are described in Section **2b2.3.**

**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*)

The frequency and distribution of missing data are described in Section **2b3.3**. We did not perform statistical analyses of missing data.

The frequency and distribution of missing data are described in Section **2b2.3**. We did not perform statistical analyses of missing data.

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

As described in Section **2b3.3**, the removal of cases from the denominator and/or numerator where an abstractor submits a value of “UTD” are necessary to align with clinical guidelines and enable measure calculation. Additionally, these exclusions/exceptions limit the biasing effects of missing data. Cases where a value of “UTD” affects clinical decision making are excluded from the measure. Cases where a value of “UTD” is reflective of poor documentation are included in the denominator but excepted from the numerator. This exclusion/exception approach penalizes facilities for poor documentation, but does not artificially include cases where rapid administration of tPA may not be appropriate care.

As described in Section **2b3.3**, the removal of cases from the denominator and/or numerator where an abstractor submits a value of “UTD” are necessary to align with clinical guidelines and enable measure calculation. Additionally, these exclusions/exceptions limit the biasing effects of missing data. Cases where a value of “UTD” affects clinical decision making are excluded from the measure. Cases where a value of “UTD” is reflective of poor documentation are included in the denominator but excepted from the numerator. This exclusion/exception approach penalizes facilities for poor documentation, but does not artificially include cases where rapid administration of tPA may not be appropriate care.

1. Respondents were provided a table detailing the key measure exclusions. [↑](#footnote-ref-1)