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

**Measure Number** (*if previously endorsed*)**:** 1952

**Measure Title**: Time to Intravenous Thrombolytic Therapy

**Date of Submission**: 8/1/2019

**Type of Measure:**

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

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

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

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

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

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

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

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

**Previous testing**

Get With The Guidelines – Stroke (GWTG-Stroke) is a clinical data registry that collects information from hospitals and clinicians on patient demographics, acute outcomes, quality measures, and health outcomes. The registry was piloted in 2001 and nationally implemented in 2003. GWTG-Stroke currently has 2,243 hospitals participating in the program and is managed by the American Heart Association (AHA) and American Stroke Association (ASA).

**Current testing**

The data source is Registry data from the 2018 Get with The Guidelines Stroke Program.

**1.3. What are the dates of the data used in testing**?

**Previous testing**

10/1/2014 – 9/30/2015

**Current testing**

The data are for the time period January 1st, 2018 through December 31st, 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: Click here to describe | other: Click here to describe |

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

**Previous testing**

The total number of hospitals reporting on this measure is 841. Of those, 672 hospitals had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. For this measure, 79.9 percent of hospitals are included in the analysis, and the average number of quality reporting events is 23.9 for a total of 16,100 events. The range of quality reporting events for 672 hospitals included is from 138 to 10. The average number of quality reporting events for the remaining 20.1 percent of hospitals who aren’t included is 6.03.

**Current testing**

We received data from 2,063 hospitals reporting on this measure through the registry for the AHA/ASA Get with the Guidelines Stroke Program during the period between 1/1/2018-12/31/2018. This data set reflects information at the hospital level and our analysis of the data as a whole is reflected throughout this submission. Of those 2,063 hospitals, 1,619 hospitals had at least one patient who qualified for the measure, after accounting for exclusions and exceptions, for a total of 33,836 eligible patients. The average number of eligible patients is 21 for the 1,619 hospitals. The range of eligible patients for 1,619 hospitals is from 1 to 171.

The AHA/ASA Get with the Guidelines Stroke Program is exclusively an inpatient registry and the dataset is comprised of inpatient/hospital and emergency department services data.

**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*)

**Previous testing**

There were 16,100 patients included in this testing and analysis. These were the patients that were associated with hospitals who had 10 or more patients eligible for this measure.

**Current testing**

There were 33,836 patients included in this reliability testing and analysis. These were the patients that were associated with hospitals who had at least one eligible patient in the year under study.

**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**.

**Previous testing**

The same data sample was used for reliability testing and exclusions analysis.

**Current testing**

The same data samples were used for reliability testing and exclusions/exceptions analysis.

**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.

**Previous testing**

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

**Current testing**

We analyzed performance based on a number of variables, including age, gender, and race/ethnicity. Performance rates for these variables are provided below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AGE** | | |
|  | **<65** | **65-79** | **80+** |
| **Performance Mean** | 0.83 | 0.85 | 0.86 |

|  |  |  |
| --- | --- | --- |
|  | **Gender** | |
|  | **Male** | **Female** |
| **Performance Mean** | 0.85 | 0.84 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Race/Ethnicity** | | | | | |
|  | **Hispanic** | **Black or African American** | **American Indian or Alaska Native** | **Asian** | **White** | **Native Hawaiian or Other Pacific Islander** |
| **Performance Mean** | 0.84 | 0.84 | 0.73 | 0.88 | 0.85 | 0.81 |

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

**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*)

**Previous testing**

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in hospital performance. Reliability at the level of the specific hospital is given by:

Reliability = Variance (hospital-to-hospital) / [Variance (hospital-to-hospital ) + Variance (hospital-specific-error]

Reliability is the ratio of the hospital-to-hospital variance divided by the sum of the hospital-to-hospital variance plus the error variance specific to a hospital. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in hospital performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the hospital performance score is a binomial random variable conditional on the hospital’s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per hospital.

**Current testing**

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in site performance and the noise is the total variability in measured performance. Reliability at the level of the specific site is given by:

Reliability = Variance (site-to-site) / [Variance (site-to-site) + Variance (site-specific-error]

Reliability is the ratio of the site-to-site variance divided by the sum of the site-to-site variance plus the error variance specific to a site. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in site performance.

Variance (site-to-site) = alpha\*beta/ ((alpha + beta + 1) \* (alpha + beta) ^2)

Variance (site-specific-error) = p(1-p)/n

Where p is the passing rate for a site and n is the number of patients for that site

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the site performance score is a binomial random variable conditional on the site’s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

For this analysis Alpha = 4.6568 and Beta = 1.1956. These parameters are used to calculate the variance (site-specific-error) which is approximately equal to 0.02. Reliability is then calculated for each site using this value and the variance (site-to-site). Average reliability is reported by averaging reliability for each site with at least 1 patient for the measure.

A reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. 1

1. Adams JL, Mehrotra A, McGlynn EA, Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical\_reports/TR863. (Accessed on February 24, 2012.)

**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*)  
**Previous testing**

This measure has 0.63 reliability when evaluated at the minimum level of quality reporting events and 0.81 reliability at the average number of quality events.

**Current testing**

The average reliability for hospitals with at least one eligible patient is 0.76. We also report the average reliability at each decile of the sample shown in the table below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reliability Statistics** | | | | | | | | | | | | | | |
| **Denominator (Patients)** | **Mean** | **SD** | **Min** | **Max** | **Decile** | | | | | | | | | |
|  |  |  |  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1+ | 0.76 | 0.24 | 0.16 | 1.00 | 0.39 | 0.54 | 0.66 | 0.75 | 0.83 | 0.90 | 0.95 | 1.00 | 1.00 | 1.00 |

**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?*)  
**Current testing**

This measure has acceptable reliability and increases as the denominator size increases.

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

**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)*

**Previous testing**

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 5= Strongly Agree

**Current testing**

**Validity testing method**

STK 04 Thrombolytic Therapy (NQF 0437) was chosen as a suitable candidate for correlation analysis due to the similarities in patient population and domain. We hypothesized higher hospital performance on time to thrombolytic therapy (i.e. the percent treated with alteplase for acute ischemic stroke within 60 minutes of hospital arrival) would be correlated with higher hospital performance on STK 04 (i.e. the percent of patients with acute ischemic stroke who arrive within 2 hours that are treated with alteplase within 3 hours). The rationale for this hypothesis is that hospitals with longer time to thrombolytic therapy might not be able to treat as many patients within 3 hours, because they might struggle to start alteplase before 3 hours for patients arriving near the 2-hour mark.

Hospitals included in the analysis had at least one patient in the denominator after exclusions and exceptions were removed. Datasets were reviewed to identify shared hospitals based on the hospital identifier. Comparing performance scores of those shared Hospital IDs, the empirical analysis uses regression with Time to Intravenous Thrombolytic Therapy (NQF 1952) as the outcome and STK 04 Thrombolytic Therapy (NQF 0437) as the predictor. Results identify the multiple R value (the correlation coefficient) and P-value of the regression variables to assess the association between performance scores of these shared Hospital IDs.

We use the following guidance to describe correlation1:

|  |  |
| --- | --- |
| Correlation | Interpretation |
| 0.80 – 1.00 | Very Strong |
| 0.60 – 0.79 | Strong |
| 0.40 – 0.59 | Moderate |
| 0.20 - 0.39 | Weak |
| 0 – 0.19 | Very Weak |

1. “11. Correlation and Regression.” *The BMJ*, 21 March 2019, <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression/> .

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

The expert panel included 20 members. Panel members were comprised of experts from the AHA Council on Stroke 2015-2016 Leadership Committee. The list of expert panel members is as follows:

Mat Reeves, BVSc, PhD, FAHA

Mai Nguyen-Huynh, MD, MAS

Judith Lichtman, PhD, MPH, FAHA

Edward Jauch, MD, MS, FAHA, FACEP

Jennifer Majersik, MD, MS

Kevin Sheth, MD, FAHA

Phillip Scott, MD

Walter N. Kernan, MD

Brett Cucchiara, MD, FAHA

Mary Ann Bauman, MD

Claranne Mathiesen, MSN, RN, CNRN, SCRN

Karen Furie, MD, MPH, FAHA

Salvador Cruz-Flores, MD, MPH, FAHA, FACP

Alejandro Rabinstein, MD, FAHA

Colin Derdeyn, MD, FAHA

N. Jennifer Klinedinst PhD, PH, RN, FAHA

Jose Romano, MD, FAHA, FAAN

Barbara Lutz, PhD, RN, CRRN, FAHA, FAAN

Pooja Khatri, MD, MSc, FAHA

Kyra Becker, MD

**Current Testing**

Data from the AHA/ASA 2018 Get with the Guidelines Stroke Program were used to perform the correlation analysis for this measure. Data comes from the Registry version of Time to Intravenous Thrombolytic Therapy (NQF 1952) and STK 04 Thrombolytic Therapy (NQF 0437).

Time to Intravenous Thrombolytic Therapy (NQF 1952) was positively correlated with STK 04 Thrombolytic Therapy (NQF 0437).

**NQF #0437**

Coefficient of correlation = 0.43

Alpha level = 0.05

P-value = < 0.001

Number of shared hospitals based on Hospital identifier = 1,612

**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?*)

**Previous testing**

The results of the expert panel rating of the validity statement were as follows: N = 20; Mean rating = 4.2 and 85% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 1 response (Strongly Disagree)

2 – 0 responses

3 – 2 responses (Neither Agree nor Disagree)

4 – 8 responses

5 – 9 responses (Strongly Agree)

**Current Testing**

Time to Intravenous Thrombolytic Therapy has a moderate positive correlation with STK 04 Thrombolytic Therapy. The correlation is highly statistically significant. With a coefficient of correlation of 0.43, the correlation is moderate, significant, and confirms our hypothesis. The moderate positive correlation with STK 04 Thrombolytic Therapy demonstrates the criterion validity of the measure. The strength of the correlation is within our expectations. We did not anticipate higher correlation because there are multiple factors related to identifying eligibility, and treating all eligible patients, that are independent of the time needed to treat.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#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*)  
**Previous testing**

**Exclusions include:**

* Age < 18 years;
* Stroke occurred after hospital arrival (in ED/Obs/inpatient);
* Patients whose date/time of ED arrival and/or date/time of thrombolytic administration are blank, not documented, or N/A;
* Patients with a negative calculated time difference;
* Patients with a Date Last Known Well, but no time Last Known Well just MM/DD/YYYY;
* Patients that receive tPA greater than 4.5 hours after Last Known Well Patients transferred from outside hospital;
* And Clinical Trial.

**Exceptions include:**

* Documented eligibility or medical reason for delay in treatment.

Exclusions and exceptions were analyzed for frequency across providers.

**Current Testing**

Exclusions include:

* Age < 18 years
* Stroke occurred after hospital arrival (in ED/Obs/inpatient)

• Patients whose date/time of ED arrival and/or date/time of IV alteplase administration is blank, unknown, or MM/DD/YYYY only

* Patients with a negative calculated time difference
* Patients with a Date Last Known Well, but no time Last Known Well
* Patients that receive IV alteplase greater than 4.5 hours after Last Known Well
* Patients who received IV alteplase at an outside hospital or by EMS/Mobile Stroke Unit
* Clinical Trial

Exceptions include:

* Patients who received IV alteplase greater than 60 minutes after arrival and have a documented Eligibility or Medical Reason for delay in treatment

Exceptions and exclusions were analyzed for frequency across hospitals and deciles of exceptions were reported.

**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*)

**Previous testing**

**Exclusions Analysis:**

Amongst the 672 hospitals with the minimum (10) number of quality reporting events, there were a total of 1,950 exclusions reported. The average number of exclusions per hospital in this sample is 2.90. The overall exclusion rate is 10.8%. The range of exclusion rates for hospitals included 47% to 0%.

**Exceptions Analysis:**

Amongst the 672 hospitals with the minimum (10) number of quality reporting events, there were a total of 3,581 exceptions reported. The average number of exceptions per hospital in this sample is 5.32. The overall exception rate is 18.2%. The range of exclusion rates for hospitals included 57% to 0%.

**Current Testing**

Amongst the 1,619 included hospitals, there were a total of 12,379 exceptions and exclusions reported. The average number of exceptions and exclusions per hospital in this sample is 7.65. The proportion of exceptions to patients is 0.37. Exception deciles illustrate the spread of exceptions amongst hospitals. According to the results, 50% of hospitals had 5 or fewer exceptions and exclusions across eligible patients for the year under study.

|  |  |
| --- | --- |
| **Decile** | **Exceptions**  **+ Exclusions** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 7 |
| 7 | 9 |
| 8 | 12 |
| 9 | 17 |
| 10 | 72 |

**Exclusions:**

LOS: Length of Stay >120 days.

CLINICAL: Clinical Trials.

SYMPLOC: In-hospital Strokes.

IVTPAOUTSIDE: IV alteplase at an outside hospital or EMS / Mobile Stroke Unit

ARRTIME: Arrival Time missing.

TPATIME: IV alteplase Time missing.

NEGTIME: IV alteplase Date/Time before Arrival Date/Time.

SYMPONSET: Symptom on set >4.5 hours.

**Exceptions:**

IVTPADELAYE: # of Eligibility Reason for delay cases.

IVTPADELAYM: # of Medical Reason for delay cases.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Exclusions** | | | | | | | | |
|  | **LOS** | **CLINICAL** | **SYMPLOC** | **IVTPAOUTSIDE** | **ARRTIME** | **TPATIME** | **NEGTIME** | **SYMPONSET** |
| **Frequency** | 4 | 162 | 2274 | 742 | 82 | 273 | 291 | 1261 |

|  |  |  |
| --- | --- | --- |
| **Exceptions** | | |
|  | **IVTPADELAYE** | **IVTPADELAYM** |
| **Frequency** | 4910 | 3588 |

**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*)  
**Previous testing**

Exclusions arise when patients who are included in the initial patient or eligible population for a measure do not meet the denominator criteria specific to the intervention required by the numerator. Exclusions are absolute and apply to all patients and, therefore, are not part of clinical judgment within a measure. Exclusions, including applicable value sets, are included in the measure specifications.

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons.

Without these being removed, the performance rate would not accurately reflect the true performance of each facility, which would result in an increase in performance failures and false negatives.

AHA/ASA recommends that physicians document the specific reasons for exception in patients’ medical records for purposes of optimal patient management and audit-readiness. AHA/ASA also advocates for the systematic review and analysis of each physician’s exceptions data to identify practice patterns and opportunities for quality improvement.

**Current Testing**

The AHA/ASA follows the PCPI methodology in distinguishing between denominator exceptions and denominator exclusions.

Denominator exclusions arise when the clinical action indicated in the numerator is not appropriate for a particular group of patients who otherwise meet the denominator criteria. These are absolute and would be removed from the denominator of a measure in order to determine the eligible population. Exclusions are included in the measure specifications.

Denominator exceptions are used to remove a patient from the denominator when the patient does not receive the action required in the numerator AND that action would not be appropriate due to a patient-specific reason(s). The patient would otherwise meet the denominator criteria. Exceptions are not absolute and are based on clinical judgment or individual patient characteristics or preferences. The PCPI methodology includes two categories of exceptions for which a patient may be removed from the denominator of an individual measure: 1) medical OR 2) patient or non-medical reasons. These exception categories are not uniformly relevant across all measures. The denominator exception language may include specific examples of instances that may constitute an exception, which are intended to serve as a guide to hospitals. For measure 1952, Time to Intravenous Thrombolytic Therapy, the exception is patients who received IV alteplase greater than 60 minutes after arrival and have a documented Eligibility or Medical Reason for delay in treatment. For example, Eligibility reasons include social/religious, initial refusal, and care-team unable to determine eligibility. Medical reasons include hypertension requiring aggressive control with IV medications, further diagnostic evaluation to confirm stroke for patients with hypoglycemia (blood glucose < 50), seizures, or major metabolic disorders, and management of concomitant emergent/acute conditions such as cardiopulmonary arrest, respiratory failure (requiring intubation).

Although this methodology does not require the external reporting of more detailed exception data, the AHA/ASA recommends that hospitals document the specific reasons for exception in patients’ medical records for purposes of optimal patient management and audit-readiness. The AHA/ASA also advocates for the systematic review and analysis of each hospital’s exceptions data to identify practice patterns and opportunities for quality improvement.

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

**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***](#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.**

**Previous testing**

Not applicable

**Current Testing**

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**.

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

*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*)**:   
Previous testing**

Not applicable

**Current Testing**

Not applicable

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

Not applicable

**Current Testing**

Not applicable

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

Not applicable

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

Not applicable

**Current Testing**

Not applicable

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

**Previous testing**

Not applicable

**Current Testing**

Not applicable

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
**Previous testing**

Measures of central tendency, variability, and dispersion were calculated.

**Current Testing**

Measures of central tendency, variability, and dispersion were calculated.

**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*)  
**Previous testing**

Based on the sample of 672 included hospitals, the mean performance rate is 0.70, the median performance rate is 0.73 and the mode is 1.0. The standard deviation is 0.22. The range of the performance rate is 1.0, with a minimum rate of 0.0 and a maximum rate of 1.00. The interquartile range is 0.32 (0.56 – 0.88).

**Current Testing**

Based on the sample of 1,619 included hospitals, the mean performance rate is 0.76, the median performance rate is 0.84 and the mode is 1.0. The standard deviation is 0.26. The range of the performance rate is 1.00, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.28 (0.95–0.67). Deciles are provided in the table below:

|  |  |
| --- | --- |
| **Decile** | **Performance** |
| 1 | 0.38 |
| 2 | 0.60 |
| 3 | 0.71 |
| 4 | 0.78 |
| 5 | 0.84 |
| 6 | 0.89 |
| 7 | 0.93 |
| 8 | 0.98 |
| 9 | 1.00 |
| 10 | 1.00 |

**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?*)  
**Previous testing**

The range of performance from 0.00 to 1.00 suggests there’s clinically meaningful variation across physicians’ performance.

**Current Testing**

The range of performance from 1.00 to 0.00 suggests there’s clinically meaningful variation across hospitals’’ performance. Outliers are considered to be values less than quartile 1 (0.67) or greater than quartile 3 (0.95) by more than 1.5 the IQR (0.28) and there 99 outliers in the data set.

|  |  |
| --- | --- |
| **Quartile** | **Performance** |
| 1 | 0.67 |
| 2 | 0.84 |
| 3 | 0.95 |
| 4 | 1.00 |

Excluding those outliers, the range of performance is 0.42 to 1.00 with a mean of 0.81, median of 0.86, and standard deviation less than 0.00. Looking at the performance percentiles without outliers, 50% of the data falls at or below a performance score of 0.86 which demonstrates additional meaningful variation across providers’ performance. See table below for performance percentiles with outliers excluded:

|  |  |
| --- | --- |
| **Decile** | **Performance** |
| 1 | 0.50 |
| 2 | 0.67 |
| 3 | 0.75 |
| 4 | 0.80 |
| 5 | 0.86 |
| 6 | 0.90 |
| 7 | 0.94 |
| 8 | 0.99 |
| 9 | 1.00 |
| 10 | 1.00 |

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

**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*)

**Previous testing**

This test was not performed for this measure.

**Current Testing**

Not applicable

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

This test was not performed for this measure.

**Current Testing**

Not applicable

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

**Previous testing**

This test was not performed for this measure.

**Current Testing**

Not applicable

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

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

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

Data are not available to complete this testing.

**Current Testing**

The registry dataset provided to us by the 2018 AHA/ASA Get with the Guidelines Program was examined and tested for missing data prior to sending. The procedure is described as the following: if data uses covariates in statistical analyses, they do imputation if the variable is missing > 15% of the time. In order to enhance data quality, they also exclude hospitals that have missing medical history in more than 25% of their submitted cases.

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

Data are not available to complete this testing.

**Current Testing**

Arrival time missing and IV alteplase time missing were variables included as measure exclusions within the specifications. There were approximately 82 missing cases of arrival time and 273 missing cases of IV alteplase time missing out of the 12,379 exceptions and exclusions reported across 1,619 hospitals.

**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*)

**Previous testing**

Data are not available to complete this testing.

**Current Testing**

Given the small frequency of missing data there is no reason to believe that missing data biased the performance results due to systematic missing data. Additionally, the imputation and exclusion of missing variables and cases is a statistically acceptable approach to dealing with missing data.