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

**Measure Number** (*if previously endorsed*)**:** 0496

**Measure Title**: Median Time from ED Arrival to ED Departure for Discharged ED Patients

**Date of Submission**: 01/15/2014 (**2014 Submission**); 01/05/2018 (**2018 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 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. **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.***)

**2014 Submission: left column selections (in black).**

**2018 Submission: right column selections (in red).**

|  |  |
| --- | --- |
| **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*).

**2014 Submission:** Testing information is provided at the end of this document.[[1]](#footnote-1)

**2018 Submission:** NQF 0496 (Median Time from ED Arrival to ED Departure for Discharged ED Patients) is specified using an *overall* rate, with three sub-populations (or *strata*). The eligibility criteria for each population (the *overall* rate and each stratum) is summarized below. The detailed denominator and numerator criteria are described in sections **1.2a** through **1.2d** (below)**.**

* Overall rate: The *overall* rate includes all eligible patients.
* Reporting rate: The *reporting* rate includes cases from the *overall* rate that are not included in the *psychiatric/mental health* rate or *transfer patient* rate.
* Psychiatric/mental health rate: The *psychiatric/mental health* rate includes cases from the *overall rate* for which the principal diagnosis is captured in the *psychiatric/mental health* code set.
* Transfer patient rate: The *transfer patient* rate includes cases from the *overall* rate for which the discharge code indicates that the patient was transferred to a facility that is an acute care facility for inpatient care of the general population or a facility operated by the Department of Defense or the Department of Veteran’s Affairs.

**A note about the use of the terms “stratification” and “stratum”/”strata” with respect to this measure**: “Stratum” refers to specific sub-populations of cases included in the *overall* rate for whom group-specific measures may be informative. It is widely acknowledged that throughput times for certain sub-populations are determined principally by their specific care needs, rather than facility performance. The measure recognizes two of these groups of particular importance*―*cases for patients with diagnoses related to psychiatric/mental health conditions and cases for patients who are transferred to acute care facilities. To allow for a full assessment of facility performance and permit a more accurate comparison of performance across facilities, the measure is calculated for all cases in an *overall* rate, but also as separate sub-rates of *psychiatric/mental health* and *transfer* *patient* rates, as well as a *reporting* rate that excludes these populations. Excluding cases where patients are included in the *psychiatric/mental health* and *transfer* *patient* rates from the *reporting* rate minimizes the potential for distortion of measure performance or confounding.

Calculation of the *overall* rate is based on values for all unduplicated cases included in one or more of the sub-population rates. The *reporting* rate is mutually exclusive from both the *psychiatric/mental health* rate and the *transfer patient* rate. Cases included in the *psychiatric/mental health* rate may also be included in the *transfer patient* rate, if inclusion criteria for both strata are met. The measurement value is calculated the same for all cases and is not risk-stratified for differences in case mix. A complete list of codes can be found in *NQF 0496\_Measure Code Set*.

1. Datasets used to define the sample:

* The initial patient population for the *overall* rate is identified using data abstracted for a sample of charts from emergency department (ED) encounters with at least one of the following Current Procedural Terminology (CPT) codes for evaluation and management (E/M) care: 99281, 99282, 99283, 99284, 99285, or 99291.

1. Datasets used to define the effective sample for each rate:

* The effective sample for each strata is identified using chart-abstracted data from the initial patient population; it is determined by the criteria laid out for each denominator exclusion and numerator exception (described below) and will differ from the defining criteria for the effective samples for the other two strata. Effective samples may not be mutually exclusive; patients may be included in more than one strata if all inclusion criteria are satisfied.

1. Datasets used to identify denominator exclusions:

* Separate, specific denominator exclusions apply to each of the four strata. Denominator exclusions are identified using chart-abstracted data of cases for patients included in the initial patient population. For each strata, cases are excluded from the effective sample if they meet one or more denominator exclusions.
  + ***Overall* rate** denominator exclusions:
    - *Discharge Code* equal to “[6] Expired;”
    - *Discharge Code* equal to “[7] Left Against Medical Advice/AMA;” and,
    - *Discharge Code* equal to “[8] Not Documented or Unable to Determine (UTD).”
  + ***Reporting* rate** denominator exclusions:
    - All of the exclusions for the *overall* rate, plus:
    - *Discharge Code* equal to “[4a] Acute Care Facility—General Inpatient Care;”
    - *Discharge Code* equal to “[4d] Acute Care Facility—Department of Defense or Veteran’s Administration;” and,
    - *ICD-10-CM Principal Diagnosis Code* equal to a coderelated to a psychiatric/mental health condition (refer to *NQF 0496\_Measure Code Set* for mental health ICD-10 codes).
  + ***Psychiatric/mental health* rate** denominator exclusions***:***
    - All of the exclusions for the *overall* rate.
  + ***Transfer patient* rate** denominator exclusions:
    - All of the exclusions for the *overall* rate.

1. Datasets used to identify numerator exceptions:

* Numerator exceptions are identified using chart-abstracted data of cases for patients included in the initial patient population and are the same for all strata. NQF 0496 is a continuous measure; therefore, numerator exceptions are treated as exceptions from the effective sample (rather than exceptions from the numerator). Cases are excepted from the effective sample if one or more of the following criteria are met:
  + ***Overall* rate** numerator exceptions:
    - *ED Arrival Time* equal to “UTD;”
    - *ED Departure Date* equal to “UTD;” and,
    - *ED Departure Time* equal to “UTD.”
  + ***Reporting* rate** numerator exceptions:
    - All of the numerator exceptions for the *overall* rate.
  + ***Psychiatric/mental health* rate** numerator exceptions:
    - All of the numerator exceptions for the *overall* rate.
  + ***Transfer patient* rate** numerator exceptions:
  + All of the numerator exceptions for the *overall* rate.

1. Datasets used to capture the numerator:

* NQF 0496 is a continuous measure; therefore, numerator criteria are treated as effective sample criteria; i.e. cases that are not excluded or excepted based on the above criteria, **and** meet the following numerator criteria are included in the measure strata. The initial patient population is identified using chart-abstracted data of cases for patients included in the effective sample for each strata. Effective samples are not mutually exclusive, and cases may be included in the effective sample of more than one strata if all criteria are satisfied. For each strata, cases are included in the effective sample if all of the following criteria are met:
  + ***Overall* rate**:
    - Cases do not meet any denominator exclusion criteria for the *overall* rate; and,
    - Cases do not meet any numerator exception criteria for the *overall* rate.
  + ***Reporting* rate**:
    - Cases do not meet any denominator exclusion criteria for the *reporting* rate; and,
    - Cases do not meet any numerator exception criteria for the *reporting* rate.
* ***Psychiatric/mental health* rate**: The *ICD-10-CM Principal Diagnosis Code* is equal to a code related to a psychiatric/mental health condition;
  + - Cases do not meet any denominator exclusion criteria for the *psychiatric/mental health* rate; and,
    - Cases do not meet any numerator exception criteria for the *psychiatric/mental health* rate.
  + ***Transfer patient* rate**:
    - *Discharge Code* equal to “[4a] Acute Care Facility—General Inpatient Care;”
    - *Discharge Code* equal to “[4d] Acute Care Facility—Department of Defense or Veteran’s Administration;”
    - Cases do not meet any denominator exclusion criteria for the *transfer patient* rate; and,
    - Cases do not meet any numerator exception criteria for the *transfer patient* rate.

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

**2014 Submission**: January 1, 2012—September 30, 2012

**2018 Submission**: October 01, 2015—August 30, 2016

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

**2014 Submission**: left column selections (in black).

**2018 Submission**: right column selections (in red).

|  |  |
| --- | --- |
| **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*)

**2014 Submission**: Blank

**2018 Submission**: The number of measured entities (hospital EDs) varies by testing type and measure strata; 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*)

**2014 Submission**: Blank

**2018 Submission**: The number of patients varies by testing type and strata; 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**.

**2014 Submission**: Blank

**2018 Submission:**

**Reliability testing:** Reliability testing was conducted for all four measure strata.

Data source:

1. *Denominator:* Clinical Data Warehouse (CDW), maintained by the Centers for Medicare & Medicaid Services (CMS)
2. *Numerator:* CDW
3. *Exclusions:* CDW
4. *Exceptions:* CDW

Dates:

1. *Denominator:* October 01, 2015–August 30, 2016
2. *Numerator:* October 01, 2015–August 30, 2016
3. *Exclusions:* October 01, 2015–August 30, 2016
4. *Exceptions:* October 01, 2015–August 30, 2016

Number of facilities sampled: 3,758

Number of cases in initial patient population (before exclusions and exceptions): 2,343,102

Effective sample (sample after exclusions, exceptions, and numerator criteria applied to initial patient population): See *Exhibit 1*

Level of analysis: Case

***Exhibit 1:*** *Effective Sample Patient Characteristics by Strata*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rate Description** | **Facility Count** | **Effective Sample** | **Effective Sample Case Characteristics** | | |
| *Gender  (% male)* | *Mean age [SD] (years)* | *Race  (% non-white)* |
| *Overall* rate | 3,758 | 2,287,933 | 43.9 | 39.3 [23.8] | 20.5 |
| *Reporting* rate | 3,757 | 2,204,485 | 43.5 | 38.7 [23.6] | 20.8 |
| *Psychiatric/mental health* rate | 1,623 | 4,686 | 54.6 | 39.2 [18.1] | 25.3 |
| *Transfer patient* rate | 3,515 | 79,125 | 52.4 | 55.3 [23.7] | 11.4 |

**Validity testing – *Data element validity***: Data element validity testing was conducted for all cases abstracted by Clinical Data Abstractor Center (CDAC) auditors.[[2]](#footnote-2)

Data source:

1. *Denominator:* CDW
2. *Numerator:* CDW
3. *Exclusions:* CDW
4. *Exceptions:* CDW

Dates:

1. *Denominator:* October 01, 2015–August 30, 2016
2. *Numerator:* October 01, 2015–August 30, 2016
3. *Exclusions:* October 01, 2015–August 30, 2016
4. *Exceptions:* October 01, 2015–August 30, 2016

Number of facilities sampled: 880

Number of cases sampled (before exclusions and exceptions): 13,187

Level of analysis: Case, data element

Sample patient characteristics:

* Gender (% male): 43.4
* Mean age (years): 39.9 (standard deviation: 23.5)
* Race (% non-white): 23.3

**Validity Testing** *―* ***Face validity***

Data source: Structured qualitative survey completed by the throughput expert work group (EWG) members

Date collected: November–December 2017

Number of responses: 9

Respondent characteristics: Respondents were asked to self-identify as one or more of the following categories: insurer/purchaser; payer; clinician (7); management (2); healthcare administration (5); patient or patient advocate; caregiver (1); other – policy researcher (1); other – professional association (1).

**Exclusions analysis:** Exclusion analysis testing was conducted for all cases included in the initial patient population.

Data source:

1. *Denominator:* CDW
2. *Numerator:* CDW
3. *Exclusions:* CDW
4. *Exceptions:* CDW

Dates:

1. *Denominator:* October 01, 2015–August 30, 2016
2. *Numerator:* October 01, 2015–August 30, 2016
3. *Exclusions:* October 01, 2015–August 30, 2016
4. *Exceptions:* October 01, 2015–August 30, 2016

Number of facilities sampled: 3,758

Number of cases sampled (before exclusions and exceptions): 2,343,102

Level of analysis: Case

Sample patient characteristics:

* Gender (% male): 43.9
* Mean age (years): 39.4 (standard deviation: 23.7)
* Race (% non-white): 20.6

**Risk adjustment/risk stratification:** N/A—this measure is not risk-adjusted or risk-stratified.

**Identification of statistically significant & meaningful differences in performance:** Identification of statistically significant and meaningful differences in performance used *Hospital Compare* data. This dataset reports facility-level measure scores for the *reporting* rate and does not include results for the o*verall*, *psychiatric/mental health*, or *transfer patient* rates. Therefore, the results of this section reflect an analysis of the *reporting* rate only.

Data Source: *Hospital Compare* downloadable dataset [maintained by CMS]

Dates:

1. *Denominator:* January 1, 2016–December 31, 2016
2. *Numerator:* January 1, 2016–December 31, 2016
3. *Exclusions:* January 1, 2016–December 31, 2016
4. *Exceptions:* January 1, 2016–December 31, 2017

Number of facilities: 3,737

Effective sample (denominator after exclusions): 2,134,653

Level of analysis: Facility

Effective sample characteristics: N/A—data available on *Hospital Compare* do not support this analysis.

**Missing data analysis and minimizing bias:** Missing data analysis testing was conducted for all cases included in the initial patient population.

Data source:

1. *Denominator:* CDW
2. *Numerator:* CDW
3. *Exclusions:* CDW
4. *Exceptions:* CDW

Dates:

1. *Denominator:* October 01, 2015–August 30, 2016
2. *Numerator:* October 01, 2015–August 30, 2016
3. *Exclusions:* October 01, 2015–August 30, 2016
4. *Exceptions:* October 01, 2015–August 30, 2016

Number of facilities sampled: 3,758

Number of cases in the initial patient population (before exclusions and exceptions): 2,343,102

Level of analysis: Case

Sample patient characteristics:

* Gender (% male): 43.9
* Mean age (years): 39.4 (standard deviation: 23.7)
* Race (% non-white): 20.6

**Comparability of performance scores when more than one set of specifications:** N/A—this measure only uses one set of specifications.

**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.

**2014 Submission**: N/A—This question was not included in version 6.5 of the Measure Testing Form.

**2018 Submission**:An Ordinary Least-Squares (OLS) regression model was used to estimate systematic relationships between patient-level characteristics and performance scores at the provider level, allowing for identification of performance gaps that may be linked to patient attributes or subpopulations. The following patient-level sociodemographic status (SDS) factors, derived from CDW data, were included in the model:

* Age;
* Gender;
* Race; and,
* Ethnicity.

Results of the regression tests are reported in section **1b.4** of the Measure Submission Form.

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

**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.*

**2014 Submission**: Per NQF comments received on 6/10/13, it is no longer necessary to report the results of the reliability testing when the results of the validity testing of individual data elements are reported.

**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*)

**2014 Submission**: Blank

**2018 Submission**: Reliability was calculated in accordance with the methods discussed in *Estimating Reliability and Misclassification in Physician Profiling* (2010). This approach uses a hierarchical linear model (HLM), which is appropriate for testing the reliability of continuous data that have clustered observations that may share variance as a results of common factors, such as multiple providers within one facility. HLM is a type of fixed-effects regression that allows for the calculation of the ratio of between group variance to total variance, designated the *intraclass correlation (ICC)* or *reliability score.* Thereliability score is a function of the number of facilities included in the analysis and the error variance within and across facilities; values could range from 0.00 to 1.00. A score of 0.00 attributes any measured difference to error (noise), while a score of 1.00 attributes any measured differences to a true difference in performance (signal). Generally, a minimum reliability score of 0.70 is considered sufficient to draw conclusions about groups (i.e., cases treated within the same facility). The ICC was calculated using the following equation:

Analysis was performed at the case level, accounting for clustering within facilities. Extreme values originally included in the *overall* rate were artificially censored at the 99th percentile (803 minutes).[[3]](#footnote-3) Artificially censoring outlier cases limits the biasing effects of these cases, while not rewarding facilities for poor performance. Facilities with fewer than 11 cases meeting criteria for the *overall* rate were omitted in accordance with *Hospital Compare’s* minimum case count criteria. To account for model convergence errors that resulted from the large sample size the analysis was conducted using a 25% random sample of each facility’s cases, from which reliability was estimated. To ensure results were not due to chance and to minimize sampling bias, the analysis was performed on ten separate 25% random samples. Samples were restricted to cases that met inclusion and exclusion criteria for the *overall* rate and were further restricted to cases meeting strata criteria for the *reporting*, *psychiatric/mental health*, and/or *transfer patient* rates. As a result, the sample pools are generalizable across all four measure strata.

See section **2b1.3** for validity testing of data elements.

REFERENCE:

1. Adams J.L., Mehrotra, A., & McGlynn, E.A. Estimating reliability and misclassification in physician profiling. Santa Monica, CA: RAND Corporation. 2010. Retrieved from <http://www.rand.org/pubs/technical_reports/TR863>.

**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*)

**2014 Submission**: Blank

**2018 Submission**: *Exhibit 2* summarizes the ranges of estimated performance score reliability for all four NQF 0496 strata, based on CDW data abstracted from October 2015–September 2016. The cases included in analysis represent a 25% random sample of the effective sample and were identified using the methodology described in section **2a2.2**. Reliability was measured using the ICC from an HLM model; values could range from zero to one, with higher scores reflecting greater reliability.

***Exhibit 2:*** *ICC Range by Stratum*

| **Stratum** | **Case Count (from 25% Sample)** | | **Facility Count (from 25% sample)** | | **ICC Range** | |
| --- | --- | --- | --- | --- | --- | --- |
| *Min* | *Max* | *Min* | *Max* | *Min* | *Max* |
| *Overall* rate | 572,545 | | 3,749 | | 0.869 | 0.872 |
| *Reporting* rate | 551,330 | 551,836 | 3,745 | 3,748 | 0.859 | 0.866 |
| *Psychiatric/mental health* rate *[[4]](#footnote-4)* | 1,091 | 1,225 | 552 | 645 | 0.648 | 0.803 |
| *Transfer patient* rate | 19,579 | 19,996 | 2,913 | 2,962 | 0.751 | 0.792 |

**Appendix A** describes the sample size, facility count, facility variance, error variance, and ICC for the iterations of reliability score estimation summarized in *Exhibit 2*.

**REFERENCE:**

1. Bartlett, J.W. & Frost, C. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. 2008.

**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?*)

**2014 Submission**: Blank

**2018 Submission**: Calculated using an HLM model, the reliability scores of all samples and measure strata indicate that variance due to error does not contribute significantly to variation in performance scores, demonstrating strong measure reliability. The results of this test indicate that the measure is able to identify true differences in performance between 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)*

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**: The validity of the measure was assessed using quantitative analyses to evaluate data element validity and qualitative analyses to assess face validity.

**Validity testing - *Data element validity***

The validity of critical data elements was evaluated by calculating kappa statistics (for categorical data elements) or Pearson correlation coefficients (for continuous data elements). Both tests assess the level of agreement between facility abstraction and auditor (CDAC) abstraction. For this test, CDAC is considered to be an authoritative source to which data from facility abstraction are compared. The kappa and Pearson correlation coefficient test statistics measure interrater reliability and quantify the agreement between two sources for the same observation (as a percent), after controlling for agreement by chance. Test statistic values may range from 0.00 to 1.00, where a value of 0.00 indicates zero agreement between two sources and a value of 1.00 indicates complete agreement between two sources. To estimate the statistical significance associated with the test statistics, p-values can be calculated. P-values of less than 0.001 indicate very high levels of statistical significance, and suggest the results are not due to chance.

The following classification offers an interpretation of a kappa statistic (Landis & Koch, 1977); a similar interpretation is appropriate for interpretation of Pearson correlation coefficients:

**Statistic Value** **Indication**

<0 Poor agreement

0.00–0.20 Slight agreement

0.21–0.40 Fair agreement

0.41–0.60 Moderate agreement

0.61–0.80 Substantial agreement

0.81–1.00 Almost perfect agreement

The analysis approach used serial calculations of kappa test statistics or Pearson correlation coefficients at each step of the measure calculation algorithm published in CMS’s Hospital Outpatient Quality Reporting Specifications Manual (version 11.0). Cases meeting exclusion criteria at a specific step were excluded from the analyses of all future steps. For example, if a case had a value of “6”, “7”, or “8” for *Discharge Code* (thus excluding them from the effective sample), the case would not be included in any data element validity assessment for algorithm steps after *Discharge Code*. As a result, the number of cases used to calculate each test statistic test will decrease after each exclusion step in the measure algorithm.

**Validity testing** *―* ***Face validity***

Face validity of the performance score was assessed systematically through survey of the throughput EWG. Nine EWG members participated in the data collection. Respondent perspectives include clinicians, management, and healthcare administration (see section **1.7** for more details). Prior to responding to questions related to performance score face validity, EWG members were provided detailed measure specifications.

The following statements related to performance score face validity were posed to the EWG:

1. The median time from ED arrival to ED departure for patients discharged from the ED can be accurately captured using chart-abstracted data.
2. The measure successfully assesses the median time from ED arrival to ED departure for patients discharged from the ED.
3. The median time from ED arrival to ED departure for patients discharged from the ED accurately reflects the quality of care provided in a facility’s emergency department.
4. The median time from ED arrival to ED departure for patients discharged from the ED allows users (such as CMS, clinicians, hospital administrators, patients, and other stakeholders) to distinguish good performance from bad performance.
5. Do you believe that it is appropriate and clinically meaningful to abstract the time of the observation order as the departure time for the ED Departure Time data element?

Responses to statements 1 through 4 were collected using a five-point Likert scale: *strongly agree, agree, undecided, disagree, strongly disagree,* and *do not know/not applicable.* Response options for question 5 were: *yes, not sure/do not kno*w, or *no*.

Additionally, the EWG was asked to provide feedback on the appropriateness of calculating performance scores using separate measure strata: psychiatric/mental health patients, patients transferred to an acute care facility (general inpatient care), and patients transferred to an acute care facility (Department of Defense or Veteran’s Administration facility). Response options were*: keep this stratification*, *remove this stratification,* and *no not know/not applicable.*

**REFERENCE:**

Landis, J. & Koch, G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159-174. 1977.

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

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**:

**Validity testing** *―* ***Data element validity***

Results of critical data element validity testing indicate almost perfect levels of agreement between the facilities’ abstraction of critical data elements and CDAC’s abstraction of data elements for the same sample of cases. The test statistic and p-value for each critical data element is provided in *Table 3* below, as well as the effective sample size used in the calculation.

***Table 3:*** *Data Elements Validity Testing Results*

| **Data Element** | **Test Statistic (p-value)** | **Effective Sample** |
| --- | --- | --- |
| *Discharge Code* a | 1.0 (<0.001) | 13,187 |
| *Arrival Time* b | 1.0 (<0.001) | 12,410 |
| *ED Departure Date* b | 1.0 (<0.001) | 12,410 |
| *ED Departure Time* b | 1.0 (<0.001) | 12,410 |
| *Measurement Value* c | - | - |
| *ICD-10-CM Principal Diagnosis Code* a | 1.0 (<0.001) | 12,410 |

a. The test statistic to assess validity for this data element is a Kappa score.

b. The test statistic to assess validity for this data element is a Pearson’s correlation.

c. This data element is a calculated value, not an abstracted value.

**Validity testing** *―* ***Face validity***

Results of the face validity assessment indicate that a diverse group of stakeholders believe the measure is a valid representation of facility performance. Results for each of the questions related to face validity are included in the six tables below.

1. *The median time from ED arrival to ED departure for patients discharged from the ED can be accurately captured using chart-abstracted data.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 33% | 3 |
| *Agree* | 56% | 5 |
| *Undecided* | 11% | 1 |
| *Disagree* | 0% | 0 |
| *Strongly Disagree* | 0% | 0 |
| *Do Not Know or Not Applicable* | 0% | 0 |

1. *The measure successfully assesses the median time from ED arrival to ED departure for patients discharged from the ED.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 22% | 2 |
| *Agree* | 56% | 5 |
| *Undecided* | 22% | 2 |
| *Disagree* | 0% | 0 |
| *Strongly Disagree* | 0% | 0 |
| *Do Not Know or Not Applicable* | 0% | 0 |

1. *The median time from ED arrival to ED departure for patients discharged from the ED accurately reflects the quality of care provided in a facility’s emergency department.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 0% | 0 |
| *Agree* | 67% | 6 |
| *Undecided* | 11% | 1 |
| *Disagree* | 11% | 1 |
| *Strongly Disagree* | 11% | 1 |
| *Do Not Know or Not Applicable* | 0% | 0 |

1. *The median time from ED arrival to ED departure for patients discharged from the ED allows users (such as CMS, clinicians, hospital administrators, patients, and other stakeholders) to distinguish good performance from bad performance.*

| **Response Option** | **Response Percentage** | **Response Count** |
| --- | --- | --- |
| *Strongly Agree* | 0% | 0 |
| *Agree* | 78% | 7 |
| *Undecided* | 0% | 0 |
| *Disagree* | 22% | 2 |
| *Strongly Disagree* | 0% | 0 |
| *Do Not Know or Not Applicable* | 0% | 0 |

1. *Do you believe that it is appropriate and clinically meaningful to abstract the time of the observation order as the departure time for the ED Departure Time data element?*

|  |  |  |
| --- | --- | --- |
| **Response Option** | **Response Percentage** | **Response Count** |
| *Yes* | 67% | 6 |
| *No* | 22% | 2 |
| *Not Sure or Do Not Know* | 11% | 1 |

1. *To be included in the NQF #0496 (OP-18)[[5]](#footnote-5) measure population, each patient must receive care in the emergency department. These patients are identified based on evaluation and management (E&M) codes used in the ED. From this initial patient population, certain patients are separated into additional rates beyond the publicly reported values for OP-18 (OP-18b), based on the situations listed in the table below. These patients are stratified into* psychiatric/mental health *rate* *(OP–18c) and* transfer patient *rate (OP–18d).*

|  |  |  |  |
| --- | --- | --- | --- |
| **Response Option** | **Keep this Stratification** | **Remove this Stratification** | **Do Not Know or Not Applicable** |
| *Psychiatric/mental health patients* | 100% (9) | 0% | 0% |
| *Patients transferred to an acute care facility (general inpatient care)* | 67% (6) | 0% | 33% (3) |
| *Patients transferred to an acute care Facility (Department of Defense or Veteran's Administration Facility)* | 44% (4) | 11% (1) | 44% (4) |

**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?*)

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**: Results of the quantitative and qualitative analysis are positive and support the conclusion that the measure and its calculation are valid representations of facility performance. Based on the Landis and Koch classification scale, described in Section **2b1.2**, there was almost perfect agreement between facility and auditor abstraction of data elements. All estimated kappa statistic and Pearson correlation coefficient values were equal to 1.0 and were statistically significant (Section **2b1.3**). This suggests strong validity for the critical data elements of the measure, as currently specified.

Nine EWG members, with backgrounds in healthcare administration, management, and clinical expertise in emergency medicine, pediatric emergency medicine, and clinical pharmacy, provided feedback on the face validity of NQF 0496 through an online survey. Most respondents agreed or strongly agreed that the median time from ED arrival to ED departure for patients discharged from the ED can be accurately captured using chart-abstracted data. Seven of the nine respondents agreed or strongly agreed that NQF 0496 successfully assesses the median time from ED arrival to ED departure for patients discharged from the ED and also allows users to distinguish good performance from bad performance. One respondent considers the variability in time stamping may impact the validity of the measure. Six of the nine respondents believe it is appropriate and clinically meaningful to abstract the time of the observation order as the departure time for the *ED Departure Time* data element.

The respondents generally support the performance score face validity of NQF 0496, although two respondents do not consider time between arrival and discharge to be a valid measure of quality because it may create incentives to discharge patients quickly (potentially before symptoms are treated) or to admit them, when observation would suffice. Others recognize that, although length of stay is not the only measure of quality, it is an important quality metric for assessing patient flow and efficiency, as a reflection of the provider’s care, and also of the ancillary staff and facility.

The EWG members were also asked to provide feedback on the appropriateness of measure strata. The measure recognizes two groups of particular significance—patients with psychiatric/mental health diagnoses and patients transferred to other acute care facilities. In order to remove the effects of these groups on overall facility performance scores, both are eliminated from the effective sample used to calculate the *reporting* rate, published on CMS’ *Hospital Compare* site, and reported separately for internal quality improvement purposes as the *psychiatric/mental health* and *transfer patient* rates. All of the EWG members supported keeping the *psychiatric/mental health* rate, and most of the EWG members supported the rate for patients transferred to general inpatient care at another acute care facility. Four of the EWG members supported keeping the stratum for patients transferred to a Department of Defense or Veteran’s Administration acute care facility; four did not know whether to keep or remove this stratum, with some confusion on the separation of VA hospitals from other acute hospitals. One respondent recommended that other conditions be considered for removal from the *reporting* rate, so that the measure could evaluate how patient acuity may impact the median time for those with less critical conditions.

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

**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*)

**2014 Submission**: Blank

**2018 Submission**: 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 October 2015 to September 2016 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.

**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*)

**2014 Submission**: Blank

**2018 Submission**: We examined overall frequencies and proportions of cases excluded for each exclusion/exception criterion, among all sampled cases, for 3,758 facilities. The sampled population included 2,343,102 cases where a patient had an ED encounter. Details for these analyses are described in *Table 4*.

***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 Equal to 6, 7, or 8* | X |  | 43,523 | 1.9 | 0.7 | 1.3 | 2.4 |
| *ED Arrival Time* |  | X | 604 | 0.0 | 0.0 | 0.0 | 0.0 |
| *ED Departure Date* |  | X | 30,479 | 1.3 | 0.2 | 0.9 | 2.0 |
| *ED Departure Time* |  | X | 37,721 | 1.6 | 0.4 | 1.1 | 2.3 |
| *ICD-10-CM-Principal Diagnosis Code [[6]](#footnote-6)* | X |  | 4,835 | 0.2 | 0.0 | 0.0 | 0.2 |
| *Discharge Code equal to 4a or 4d [[7]](#footnote-7)* | X |  | 79,729 | 3.4 | 0.8 | 2.3 | 5.2 |
|  |  |  |  |  |  |  |  |
| *Total Denominator Exclusions* | 3 exclusions | - | 127,608 | 5.5 | 2.8 | 4.7 | 7.5 |
| *Total Numerator Exceptions* | - | 3 exceptions | 37,972 | 1.6 | 0.4 | 1.2 | 2.3 |
|  |  |  |  |  |  |  |  |
| *Total Removed from the Denominator or Numerator* | 6 exceptions and exclusions | | 138,617 | 5.9 | 3.1 | 5.1 | 8.1 |

**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*)

**2014 Submission**: Blank

**2018 Submission**: As seen in *Table 4* (section **2b2.2** above), the frequency of exclusions/exceptions were low and varied minimally across facilities, as evidenced by the small interquartile range for each exclusion/exception tested. Despite the low frequency of each exclusion/exception, however, removal of cases where patients had a psychiatric/mental health diagnosis or were transferred to other acute care facilities were supported by the throughput EWG.

Measure exclusion and exception criteria are in alignment with clinical guidelines and also ensure that all cases included in the measure have sufficient information to calculate the performance score. After identification of cases for patients with an ED encounter, exclusion and exception criteria are applied. In the case of continuous measures, cases excepted from the numerator are excepted from the effective sample; therefore, in continuous measures, exclusion and exceptions are treated the same to ensure calculation of the measurement value is possible.

1. *Discharge Code* is a denominator exclusion criterion that is applied in two separate steps in the measure algorithm. In the first step, cases for patients where *Discharge Code* equals “[6] Expired,” “[7] Left Against Medical Advice/AMA,” or “[8] Not Documented or Unable to Determine (UTD)” are excluded from the effective sample. The second step is described below. Overall, 1.9% of cases for patients included in the sample are excluded from the effective sample based on *Discharge Code* (step one). There is minimal variability in the proportion of cases excluded based on *Discharge Code* values across facilities, with an interquartile range of 0.7% to 2.4%.
2. *Arrival Time* is a numerator exception criterion. If *Arrival Time* is equal to “UTD,” the case is excepted from the effective sample. Overall, less than 0.1% of cases for patients included in the sample have a “UTD” value for *Arrival Time*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important because a “UTD” value for this data element makes it impossible to determine the time from ED arrival to discharge.
3. *ED Departure Date* is a numerator exception criterion. If *ED Departure Date* is equal to “UTD,” the case is excepted from the effective sample. Overall, 1.3% of cases for patients included in the sample have a “UTD” value for *ED Departure Date*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important because a “UTD” value for this data element makes it impossible to determine the time from ED arrival to discharge.
4. *ED Departure Time* is a numerator exception criterion. If *ED Departure Time* is equal to “UTD,” the case is excepted from the effective sample. Overall, 1.6% of cases for patients included in the sample have a “UTD” value for *ED Departure Time*. While there is limited variability in the proportion of excepted cases across facilities, the exception remains important because a “UTD” value for this data element makes it impossible to determine the time from ED arrival to discharge.
5. *ICD-10-CM Principal Diagnosis Code* is a denominator exclusion criterion. Cases for patients where *ICD-10-CM Principal Diagnosis Code* is equal to a psychiatric/mental health condition are excluded from the effective sample for the *reporting* rate only. Overall, 0.2% of cases for patients included in the sample are excluded from the effective sample based on a psychiatric/mental health condition. There is limited variability in the proportion of cases excluded based *ICD-10-CM Principal Diagnosis Code.*
6. *Discharge Code* is a denominator exclusion criterion that is applied in two steps of the measure algorithm. Exclusion during an earlier step in measure calculation is described above. In the second step, cases for patients where *Discharge Code* is equal to “[4a] Acute Care Facility—General Inpatient Care” or “[4d] Acute Care Facility—Department of Defense or Veteran’s Administration” are excluded from the effective sample for the *reporting* rate only. Overall, 3.4% of cases for patients included in the sample are excluded from the effective sample based on *Discharge Code* (phase two). There is minimal variability in the proportion of cases excluded based on *Discharge Code* values across facilities, with an interquartile range of 0.8% to 5.2%.

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

**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: **2014 submission:** The results are stratified by reporting/non-reporting. The non-reporting group contains cases that were transferred or who had a psych diagnosis.

**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.**

**2014 Submission**:Blank

**2018 Submission**: Not applicable—No risk adjustment or risk stratification was performed.

**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**.

**2014 Submission**:Blank

**2018 Submission**:

This measure is a process measure for which we provide no risk adjustment or risk stratification. We determined risk adjustment and risk stratification were not appropriate based on the measure evidence base and the measure construct. As a process-of-care measure, timely discharge from the ED 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 effective sample.

**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?

**2014 Submission**:Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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)** Not applicable—No risk adjustment or risk stratification was performed.

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

**2014 Submission**:Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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.**

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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*)

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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*)**:**

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk stratification was performed.

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

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk stratification was performed.

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

**2014 Submission:** Blank

**2018 Submission**: Not applicable— No risk adjustment or risk stratification was performed.

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

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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*)

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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*)

**2014 Submission:** Blank

**2018 Submission**: Not applicable—No risk adjustment or risk 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)*

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**: Differences in performance scores and the mean performance score for facilities meeting public reporting requirements were tested. For the **January 1, 2016** to **December 31, 2016** data collection period, this included 3,737 facilities. 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*)

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**:

***Table 5:*** *Distribution of Facility Performance Scores*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Std. Dev. (SD)** | **Min.** | **10th Percentile** | **25th Percentile** | **Median** | **75th Percentile** | **90th Percentile** | **Max.** |
| 141.7 | 42.1 | 45 | 94 | 112 | 136 | 165 | 217 | 440 |

**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?*)

**2014 Submission**: Refer to Appendix B for the 2014 response to this question.

**2018 Submission**: The measure is able to discriminate between facilities based on their performance score and is able to detect differences in performance above and below the mean score. Facility performance scores ranged from 45 minutes to 440 minutes, with a median of 136 minutes. The mean ± standard deviation facility performance score was 141.7 minutes ± 42.1 minutes.

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

**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*)

**2014 Submission:** Blank

**2018 Submission**: Not Applicable—this measure uses only one set of specifications.

**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*)

**2014 Submission:** Blank

**2018 Submission**: Not Applicable—this measure uses only one set of specifications.

**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*)

**2014 Submission:** Blank

**2018 Submission**: Not Applicable—this measure uses only one set of specifications.

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

**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*)

**2014 Submission:** Blank

**2018 Submission**: NQF 0496 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 facilities 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, *ED Departure Time*. 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. In the case of continuous measures, cases excepted from the numerator are excepted from the effective sample; therefore, in continuous measures, exclusion and exceptions are treated the same to ensure calculation of the measurement value is possible. 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 **2b2.2** above.

**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*)

**2014 Submission:** Blank

**2018 Submission**: The frequency and distribution of missing data are described in section **2b2.2**. 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*)

**2014 Submission:** Blank

**2018 Submission**: As described in section **2b2.2**, the removal of cases from the effective samples 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. As noted in section **2b6.1**, continuous measures treat exclusions and exceptions the same, removing them from the effective sample. Overall, 37,972 cases of the 2,343,102 cases in the sample (1.6%) have “UTD” value for the three numerator exception criteria, suggesting that removal of these cases have a negligible effect on measure scores. The frequency and distribution of numerator exceptions are discussed in section **2b2.2.**

***Appendix A****: Reliability Iteration Results*

| **Measure Stratification** | **Sample Pool** | **Sample Size Meeting Stratum Criteria** | **Facility Count Meeting Stratum Criteria** | **Facility Variance** | **Error Variance** | **ICC** |
| --- | --- | --- | --- | --- | --- | --- |
| *Overall* rate | 1 | 572,454 | 3,749 | 14744.59 | 2231.09 | 0.869 |
| 2 | 572,454 | 3,749 | 14680.31 | 2199.39 | 0.870 |
| 3 | 572,454 | 3,749 | 14834.54 | 2198.29 | 0.871 |
| 4 | 572,454 | 3,749 | 14773.34 | 2200.19 | 0.870 |
| 5 | 572,454 | 3,749 | 14782.17 | 2213.29 | 0.870 |
| 6 | 572,454 | 3,749 | 14749.34 | 2160.01 | 0.872 |
| 7 | 572,454 | 3,749 | 14866.51 | 2182.05 | 0.872 |
| 8 | 572,454 | 3,749 | 14681.48 | 2176.81 | 0.871 |
| 9 | 572,454 | 3,749 | 14800.74 | 2202.04 | 0.870 |
| 10 | 572,454 | 3,749 | 14718.69 | 2175.51 | 0.871 |
| *Reporting* rate | 1 | 551,836 | 3,747 | 13998.75 | 2302.61 | 0.859 |
| 2 | 551,781 | 3,748 | 13893.86 | 2270.62 | 0.860 |
| 3 | 551,330 | 3,747 | 14039.02 | 2261.20 | 0.861 |
| 4 | 551,649 | 3,747 | 13963.56 | 2164.53 | 0.866 |
| 5 | 551,620 | 3,747 | 14009.66 | 2297.73 | 0.859 |
| 6 | 551,415 | 3,745 | 13975.28 | 2228.80 | 0.862 |
| 7 | 551,564 | 3,747 | 14085.29 | 2263.01 | 0.862 |
| 8 | 551,663 | 3,748 | 13923.31 | 2255.05 | 0.861 |
| 9 | 551,373 | 3,748 | 13999.63 | 2289.33 | 0.859 |
| 10 | 551,762 | 3,748 | 13946.69 | 2254.35 | 0.861 |
| *Psychiatric/mental health* rate | 1 | 1,091 | 552 | 51828.86 | 20636.08 | 0.715 |
| 2 | 1,194 | 640 | 55050.24 | 16032.59 | 0.774 |
| 3 | 1,225 | 645 | 47726.21 | 25884.94 | 0.648 |
| 4 | 1,206 | 638 | 51893.78 | 22645.87 | 0.696 |
| 5 | 1,180 | 621 | 52375.28 | 19566.91 | 0.728 |
| 6 | 1,178 | 633 | 55255.09 | 13594.51 | 0.803 |
| 7 | 1,118 | 626 | 48742.21 | 25790.13 | 0.654 |
| 8 | 1,162 | 619 | 50810.02 | 20506.16 | 0.712 |
| 9 | 1,205 | 626 | 51869.10 | 16801.21 | 0.755 |
| 10 | 1,176 | 627 | 52529.29 | 18431.81 | 0.740 |
| *Transfer patient* rate | 1 | 19,610 | 2,937 | 19168.39 | 6364.45 | 0.751 |
| 2 | 19,579 | 2,907 | 19300.20 | 6043.56 | 0.762 |
| 3 | 19,996 | 2,939 | 19424.10 | 5782.03 | 0.771 |
| 4 | 19,685 | 2,934 | 18630.90 | 5823.01 | 0.762 |
| 5 | 19,735 | 2,919 | 19250.39 | 5065.81 | 0.792 |
| 6 | 19,954 | 2,958 | 19141.47 | 5597.58 | 0.774 |
| 7 | 19,858 | 2,962 | 19168.28 | 5981.36 | 0.762 |
| 8 | 19,723 | 2,922 | 19389.42 | 5792.72 | 0.770 |
| 9 | 19,976 | 2,937 | 19813.33 | 5507.19 | 0.783 |
| 10 | 19,614 | 2,913 | 19649.58 | 5451.38 | 0.783 |

**Appendix B: Report from 2014 Submission**

**NQF# 0496** Median Time from ED Arrival to ED Departure for Discharged ED Patients

**Reliability Testing**

* Per NQF comments received on 6/10/13, it is no longer necessary to report the results of the reliability testing when the results of the validity testing of individual data elements are reported.

**Validity Testing**

We tested the validity at the data element level:

*Population and sample:* The measure population as reported in the QIO Clinical Data Warehouse (CDW) included 2,951,297 cases from 3,393 hospitals nationwide. These cases were abstracted by the individual hospitals or their vendors and the data were submitted to the CDW. The measure period is from January 1, 2012 to September 30, 2012. The CMS contractor in charge of the Warehouse maintenance randomly selected hospitals for validation on an annual basis. Up to 12 cases were selected from each selected hospital for each quarter. The CMS contractor randomly selected 11,525 cases out of 2,951,297 cases from the CDW during the measurement period. These 11,525 sample cases originated from 888 hospitals.

*Chart Abstraction:* Both the original dataset and the sample dataset were obtained from direct medical chart abstraction. The original population dataset was abstracted by the hospitals or their vendors. The sampled validation dataset was re-abstracted by the CMS Clinical Data Abstraction Center (CDAC) using exactly the same medical charts. CDAC is a CMS contractor center that has specialized in medical chart abstraction for the last fifteen years. The CDAC-abstracted data is considered the “gold standard” for the purpose of this analysis.

*Validity Test*: There are six critical data elements for this measure. We conducted validity testing on all six critical data elements. For each data element, we calculated the raw agreement rate between data from the hospital chart abstractor and the CDAC re-abstractor. We reported the Kappa statistic for the categorical data elements with binary Yes/No values. Kappa is a measure of inter-rater agreement that accounts for abstractors’ agreement by chance alone. It is standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e., potential systematic disagreement between the abstractors. A common scale is used to interpret Kappa statistics: 0.01–0.20 is slight agreement; 0.21– 0.40 is fair agreement; 0.41–0.60 is moderate agreement; 0.61–0.80 is substantial agreement; 0.81–0.99 is almost perfect agreement. For data elements with continuous value, such as date or time, we calculated the Intraclass Correlation Coefficient (ICC). Like the Kappa statistics, the ICC also accounts for the abstractors’ agreement by chance alone.

*Results.* Table 3 below shows that the sampled validation dataset was a fair representation of the original population. All the segments of the original population are present in the validation sample. Overall, the distributions of the patient and hospital characteristics in the sampled dataset are similar to those in the original population. Patient characteristics in the table included age, gender, and race/ethnicity. Hospital characteristics included bed size, teaching status, and urban vs. rural location.

Table 4 summarizes the results of the validity test of the six data elements. Overall, the agreement rates were high. The agreement rates for all data elements were higher than 90%. One data element, Observation Services, had a high agreement rate (98.14%) and a very low kappa (0.17). The potential reason for the discrepancy between the agreement and kappa is that observation services=Y is a very rare occurrence. The Kappa statistic is affected by the prevalence of the data of interest. For data of rare occurrence, very low values of kappa may not necessarily reflect low overall agreement. The definition of this data element (Observation Services) has recently been updated to ease its abstraction from the medical charts. The kappa statistic or ICC for all other seven data elements reflected almost perfect agreement.

**Table 3**. The distribution of patient and hospital characteristics between Sample and Population

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample | | Population | |
|  | Frequency | Percent | Frequency | Percent |
| **Patient Characteristics** |  |  |  |  |
| **Gender** |  |  |  |  |
| Male | 5,059 | 43.90 | 1,292,574 | 43.80 |
| Female | 6,466 | 56.10 | 1,658,334 | 56.19 |
| Undetermined |  |  | 389 | 0.01 |
| **Race** |  |  |  |  |
| Caucasian | 7,580 | 65.77 | 1,945,210 | 65.91 |
| African American | 1,919 | 16.65 | 503,923 | 17.07 |
| Hispanic | 1,263 | 10.96 | 302,135 | 10.24 |
| Native American | 115 | 1.00 | 24,452 | 0.83 |
| Asian | 150 | 1.30 | 39,369 | 1.33 |
| Other/UTD | 498 | 4.32 | 136,208 | 4.62 |
| **Age Category** |  |  |  |  |
| Under 65 | 9,645 | 83.69 | 2,508,558 | 85.00 |
| Age 65\_74 | 845 | 7.33 | 197,114 | 6.68 |
| Age 75\_84 | 660 | 5.73 | 153,570 | 5.20 |
| Age 85 plus | 375 | 3.25 | 92,055 | 3.12 |
| **Hospital Characteristics** |  |  |  |  |
| **Bed Size** |  |  |  |  |
| 1 - 100 | 284 | 31.98 | 1,215 | 35.81 |
| 101 - 200 | 234 | 26.35 | 791 | 23.31 |
| 201 - 300 | 129 | 14.53 | 497 | 14.65 |
| 301 - 400 | 95 | 10.70 | 339 | 9.99 |
| 401 plus | 146 | 16.44 | 551 | 16.24 |
| **Teaching Status** |  |  |  |  |
| Yes | 265 | 29.84 | 960 | 28.29 |
| No | 623 | 70.16 | 2,433 | 71.71 |
| **Location** |  |  |  |  |
| Rural | 297 | 33.45 | 1,198 | 35.31 |
| Urban | 591 | 66.55 | 2,195 | 64.69 |

**Table 4: Validity Test Summary for Measure 0496 (Q1 – Q3, 2012)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Eligible Cases (Denominator) | Number of cases in agreement | Agreement Rate (%) | Kappa Statisticsa/ICC\* |
| Discharge Status | 8,391 | 8,387 | 99.95 | n/a |
| E/M Code | 11,292 | 11,291 | 99.99 | n/a |
| Principal Diagnosis Code | 11,515 | 11,515 | 100.00 | n/a |
| ED Departure Date | 11,416 | 11,304 | 99.02 | 0.99\* |
| ED Departure Time | 11,369 | 10,338 | 90.93 | 0.99\* |
| Observation Services | 11,520 | 11,306 | 98.14 | 0.17a |

a - Kappa Statistics

\* - ICC

1. 2014 testing information is included in Appendix B. [↑](#footnote-ref-1)
2. CDAC is considered to be an authoritative source to which data from facility abstraction are compared. [↑](#footnote-ref-2)
3. The 99th percentile is based on the measure score of cases included in the *Overall* rate. [↑](#footnote-ref-3)
4. Due to the limited cases eligible for the P*sychiatric/Mental Health* rate within each sample, reliability was estimated for the all cases in the effective sample (4,686 cases; 1,623 facilities) as well. The ICC is equal to 0.700, which is within the range of ICC values estimated for the samples. [↑](#footnote-ref-4)
5. Questions in the EWG survey refer to the measure as OP-18, as the experts are familiar with the nomenclature used in the OQR program (OP-18a for *overall* rate, OP-18b for *reporting* rate, OP-18c for *psychiatric/mental health* rate, and OP-18d for *transfer patient* rate). [↑](#footnote-ref-5)
6. *ICD-10-CM Principal Diagnosis Code* equal to a code related to a psychiatric/mental health condition is a denominator exclusion for the *Reporting* rate. Please note: a value for the *ICD-10-CM Principal Diagnosis Code* data element that is for a psychiatric/mental health condition is a numerator condition captured by the *Psychiatric/Mental Health* rate. [↑](#footnote-ref-6)
7. *Discharge Code* equal to “4a” or “4d” is a denominator exclusion for the *Reporting* rate. Please note: a value for the *Discharge Code* data element that is equal to “4a” or “4d” is a numerator condition captured by the *transfer patient* rate. [↑](#footnote-ref-7)