NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0496

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

Measure Steward: Centers for Medicare and Medicaid Services

Brief Description of Measure: NQF #0496 calculates the median time from emergency department arrival to time of departure from the emergency room for patients discharged from the emergency department (ED). The measure is calculated using chart-abstracted data, on a rolling quarterly basis, and is publically reported in aggregate for one calendar year. The measure has been publically reported since 2013 as part of the ED Throughput measure set of the CMS' Hospital Outpatient Quality Reporting (HOQR) Program.

Developer Rationale: Empirical evidence demonstrates that ED throughput is an indicator of hospital quality of care, and shows that shorter lengths of stay in the ED lead to improved clinical outcomes. Significant ED overcrowding has numerous downstream effects, including prolonged patient waiting times, increased suffering for those who wait, rushed and unpleasant treatment environments, and potentially poor patient outcomes (Gardner, 2018). Quality improvement efforts aimed at reducing ED overcrowding and length of stay have been associated with an increase in ED patient volume, decrease in number of patients who leave without being seen, reduction in costs, and increase in patient satisfaction (Bucci, 2016; Chang, 2017; Zocchi, 2015). An analysis of data from 2,619 hospitals support that reducing the time patients remain in the ED is associated with increased patient satisfaction and a decreased chance that patients will leave before being seen (Chang, 2017). Recent guidelines and peer-reviewed studies also demonstrate the need for dedicated emergency mental health services, providing evidence that the clinical needs for these patients substantively differ from the non-psychiatric population (Nazarian, 2017; Lester, 2018).

REFERENCES:

1) Bucci, S., A. G. de Belvis, S. Marventano, A. C. De Leva, M. Tanzariello, M. L. Specchia, W. Ricciardi and F. Franceschi. (2016). Emergency department crowding and hospital bed shortage: Is Lean a smart answer? A systematic review. Eur Rev Med Pharmacol Sci, 20(20), 4209-4219.
2) Chang, A. M., A. Lin, R. Fu, K. J. McConnell and B. Sun. (2017). Associations of Emergency Department Length of Stay With Publicly Reported Quality-of-care Measures. Acad Emerg Med, 24(2), 246-250.
3) Gardner, R. M., N. A. Friedman, M. Carlson, T. S. Bradham and T. W. Barrett. (2017). Impact of revised triage to improve throughput in an ED with limited traditional fast track population. Am J Emerg Med., 36(1), 124-127.
4) Lester, N. A., L. R. Thompson, K. Herget, J. A. Stephens, J. V. Campo, E. J. Adkins, T. E. Terndrup and S. Moffatt-Bruce. (2017). CALM Interventions: Behavioral Health Crisis Assessment, Linkage, and Management Improve Patient Care. Am J Med Qual., 33(1), 65-71.
5) Nazarian DJ, Broder JS, Thiessen ME, Wilson MP, Zun LS, Brown MD, American College of Emergency Physicians. Clinical policy: critical issues in the diagnosis and management of the adult psychiatric patients in the emergency department. Ann Emerg Med. 2017 Apr; 69(4):480-98. Guideline available at: http://www.annemergmed.com/article/S0196-0644(17)30070-7/pdf.
6) Zocchi, M. S., M. S. McClelland, and J. M. Pines. Increasing Throughput: Results From A 42-Hospital Collaborative To Improve Emergency Department Flow. The Joint Commission Journal on Quality and Patient Safety, 2015, 41(12):532–542.

Numerator Statement: Continuous Variable Statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.
Denominator Statement: This measure is reported as a continuous variable statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.
Denominator Exclusions: Patients who expired in the emergency department, left against medical advice (AMA), or whose discharge was not documented or unable to be determined (UTD) are excluded from the target population.

Measure Type: Process
Data Source: Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Sep 09, 2014

Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

The developer submitted empirical evidence (literature review) but without systematic review and grading of the evidence.

Evidence Summary or **Summary of prior review in [year]**

This measure calculates the median time from emergency department (ED) arrival to time of departure from the emergency room for patients discharged from the emergency department. Facilities that report a high median time from arrival to departure may be more likely experience ED crowding, which is associated with unfavorable health outcomes, including longer hospital stays, increased costs, and higher mortality rates (Sun et al., 2013). Studies indicate that by reducing ED throughput times, facilities can increase ED patient volume, decrease the number of patients who leave without being seen, reduce costs, and increase patient satisfaction (Bucci et al., 2016; Chang et al., 2017; Zocchi et al., 2015).

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☒ **The developer provided updated evidence for this measure:**

Updates:

The measure developer conducted a review of clinical practice guidelines, peer-reviewed literature, and related policy during the NQF #0496's annual literature review to identify additional evidence and/or new studies that support the measure's intent. The measure developer identified relevant peer-reviewed publications by searching the PubMed MEDLINE database for evidence made available from January 1, 2013 to September 30, 2017.

The following studies and findings were identified through the literature search.

- Gardner et al. note that ED throughput is a meaningful indicator of hospital quality of care, and validates that shorter ED lengths of stay lead to improved clinical outcomes. In this study a revised triage process on ED throughput was associated with improvements in several ED throughput metrics and a reduction in patients left without being seen (2018).
- Mullins et al. studied data from *Hospital Compare*, which use the *Reporting Rate* strata for NQF #0496; the research team concluded that there is widespread variation in performance across the United States and that ED crowding is linked to inpatient quality outcomes (2014).
- Chang et al. conducted an analysis of data from 2,619 hospitals, showing that reducing ED length of stay is associated with increased patient satisfaction and decreased likelihood that a patient will leave before a medical professional sees him or her (2017).
- Authors of multiple studies (Melton et al. 2016; Allaudeen et al. 2017; Bucci et al. 2016) describe quality improvement and Lean-based interventions, which aim to improve ED throughput time and show that ED crowding and timely throughput remain high-priority issues for hospitals .
- A 2017 guideline prepared by the American College of Emergency Physicians (ACEP) justifies the separate measurement of patients for mental health and psychiatric services (captured in the *Psychiatric/Mental Health Rate* strata), based on evidence that the clinical needs for these patients substantively differ from those patients seeking non-psychiatric treatment (Nazarian et al. 2017).

Questions for the Committee:

- o *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- o *If the Committee believes there is a need to discuss the evidence for this measure:*
  - ▪ *What is the relationship of this measure to patient outcomes?*
  - ▪ *How strong is the evidence for this relationship?*
  - ▪ *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

Process measure (Box 3) → Literature review not including grading of body of evidence (Box 7) → Empirical evidence submitted (Box 8) → Includes whole body of evidence (Box 9) → Benefits outweigh undesirable effects→ Moderate

The highest possible rating is Moderate.

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Analysis of facility-level data from the Hospital Compare downloadable files indicates that there is variation in the median time from ED arrival to time of departure from the emergency room.

- During the January 2014 to December 2014 data collection periods, median facility-level throughput times ranged from 46 minutes to 424 minutes, with a median of 140 minutes.
- During the January 2016 to December 2016 data collection periods, median facility-level throughput times ranged from 45 minutes to 440 minutes, with a median of 136 minutes.
- When aggregating findings from both data collection periods, the median value of median time from emergency department arrival to time of departure from the emergency room decreased 2.9% (-4 minutes).

- During the January 2014 to December 2016 data collection periods, there is documentation of substantial variation in facility performance. The interquartile range has been consistently wide, ranging from 112 minutes to 165 minutes. Additionally, the maximum time for ED discharge increased between 2014 and 2016. While median performance is improving, there is an ongoing opportunity for improvement in performance at the facility level.

Disparities:
A study using data submitted to the Clinical Data Warehouse (CDW) between October 01, 2015 and August 30, 2016 examined patient and facility characteristics on ED throughput time.

- Primary results from the regression were related to patient demographics.
  - ED throughput time was significantly longer for patients in the 18–30 (ß= 31.4 minutes, p<0.001), 30–40 (ß= 41.1 minutes, p<0.001), 40–50 (ß= 53.2 minutes, p<0.001), 60–70 (ß= 70.1 minutes, p<0.001), 70–80 (ß= 77.3 minutes, p<0.001), 80–90 (ß= 84.9 minutes, p<0.001), and over 90 (ß= 91.6 minutes, p<0.001) age groups, as compared to those patients less than 18 years old.
  - There was a significantly longer ED throughput times for female patients, as compared to male patients (ß= 6.1 minutes, p< 0.001).
  - When compared to white patients, there was a significantly longer ED throughput time for Asian patients (ß= 10.3 minutes, p< 0.001); Hispanic patients also experienced longer ED throughput times, as compared to the non-Hispanic peers (ß= 12.0 minutes, p<0.001).
- ED throughput times also varied by the characteristics of the facility from which the patient was discharged.
  - When compared to patients discharged from facilities with fewer than 50 beds (a proxy for facility size), there was a significantly longer ED throughput time for patients discharged from facilities with 51–100 beds (ß= 10.7 minutes, p=0.004), 101–250 beds (ß= 35.8 minutes, p <0.001), 251–500 beds (ß= 53.8 minutes, p <0.001), and more than 500 beds (ß= 64.7 minutes, p< 0.001).
  - Urbanicity also impacted ED throughput times, with a significantly higher time for patients discharged from an urban hospital, as compared to those discharged from a rural hospital (ß= 6.5 minutes, p< 0.001).
  - When compared to patients discharged from a non-teaching facility, there was a significantly longer ED throughput time for patients discharged from a major teaching facility (ß= 54.7 minutes, p< 0.001).

Questions for the Committee:

o *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus:

- The evidence is sufficient but it is a bit disappointing that after using this measure for so many years we still do not have graded evidence.
- There appears to be sufficient evidence that supports the measure's intent to measure ED throughput as indicator of hospital quality of care. I am not aware of any new studies/information not cited that changes the evidence base for this measure.
- The measure developers assert that median time from ED admission to ED discharge indicates hospital quality of care, but cite articles that focus on patient satisfaction. I would like to see evidence that shows how amount of time patient spends in ED is related to clinical quality measures (ex. prescription of antibiotics for viral URI's, accurate diagnosis of DVT/PE and MI, hospital length of stay, mortality, hospital and ED readmission rates).
- This is a chart abstraction measure- has not changed. It would be interesting to see some more information on variability of results based on institution type and more on demographics of the patients seen. Although studies support this measure as a quality metric, things really haven't changed over the years reported- this appears to

be just as much or even more a patient experience of care metric rather than a performance/quality of care metric.

- There is evidence provided to show a relationship to improved patient experience/satisfaction and decrease in number of patients who leave without being seen, reduction in costs. No systematic review.
- Literature review since the previous measure approval shows that multiple studies have supported an association between this process measure and several outcomes including patient satisfaction.

1b. Performance Gap:

- The data on disparities is compelling and the gap is apparent.
- It is well documented that there is a gap in care that warrants measurement. There is data on the measure by population subgroups that does demonstrate potential disparities in care.
- The variation in ED times by hospital size, patient ethnicity, gender, and age is striking. This information causes me to suspect disparities in care. I think this would be useful performance measure in that it would provoke further research in to the causes. Setting a standard for median ED time would be very difficult as many factors could legitimately influence a facility's performance.
- Would like to see more disparities data.
- There is variation across EDs in performance on the measure. Measure developer shows differences by subgroups (female/male, Asian/non-Asian), and younger patients. Younger patients may have less urgent conditions to be treated, so this is not surprising result.

This measure is reported in Hospital Compare. There remains wide variability in performance across institutions in both 2014 and 2016 data. Little improvement has occurred nationally. Significant differences in median ED throughput time were observed across age, gender, ethnicity and type of institution. Larger facilities, those in urban settings and major teaching facilities have longer ED throughput times.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.
2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.
Validity
2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.
2b2-2b6. Potential threats to validity should be assessed/addressed.
Composite measures only:
2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**
Evaluators: Joseph Kunisch, Jack Needleman, and Christie Tiegland

Evaluation of Reliability and Validity (and composite construction, if applicable): Evaluation A, Evaluation B, Evaluation C

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications:

- No concerns. Agree with moderate ratings.
- The lack of adjustment for risk/case-mix adjustment and SDS factors is problematic with this measure, considering the longstanding observation that large teaching, urban hospitals have struggled with ED throughput which may be in part due to the complexity of the cases they receive through the ED and lack of community access to alternative care options outside of the ED. While hospitals should always seek to improve ED throughput, I question whether raw comparisons of all hospitals on ED throughput time are helpful and aid improvement.
- I have no concerns about reliability or validity, except that interpretation of the meaning and significance of median ED time will face similar challenges to interpretation of hospital mortality rates.
- I was not clear if admissions from the ED to inpatient status or observation are included or excluded.
- Evidence here is mixed/moderate. Kappa agreement is strange (1.0) per reviewer's comment. Not necessary to risk adjust, but model results were not displayed.
- All reviewers agreed that specifications were reliable

2a2. Reliability - Testing:

- Multiple Committee Members had no concerns regarding Criterion 2a2. Reliability.
- Limited testing done. Kappa statistic is suspicious. Not clear it discriminates performance between EDs as those data not shown.

2b1. Validity -Testing: 2b4. Meaningful Differences: 2b5. Comparability of performance scores: 2b6. Missing data/no response:

- Multiple Committee Members had no concerns regarding Criterion 2b1. Validity -Testing
- See comments about risk-adjustment and whether the comparative analysis of raw ED throughput times identifies meaningful differences in quality or in reality meaningful differences case-mix and communities served.
- Did not see results of regression models (per reviewer comments)
- Although some reviewers had specific issues/questions, all three rated overall validity as moderate.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: 2b3. Risk Adjustment:

- No concerns
- The lack of appropriate risk-adjustment is a concern.
- I do not see that any risk adjustment was performed. I would want to look more data from use of this measure before I decide if risk adjustment is needed. Based on the aforementioned variation in median ED time, this measure may be most helpful by exposing variation and promoting analysis of the root causes. Risk adjusting may obscure important variation that needs further analysis.
- Exclusions appear consistent. Again, not clear how eventual admissions to the hospital or observation unit are handled.
- Handling psychiatric patients and those transferred to acute facilities as separate strata appears sufficient for risk adjustment.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer stated that for clinical measures, the required data elements are routinely generated/collected during provision of care (e.g., blood pressure, lab value, diagnosis, medication order, depression score). Also, the data is abstracted from a record by another individual than the individual who obtained the original information (e.g., chart abstraction for quality measure/registry).
- All data elements in the electronic health records are in defined fields from a combination of electronic sources.
- Feedback on this measure were provided by nine expert work group member through an online survey. Expert member had backgrounds in healthcare administration, management, and clinical expertise in emergency medicine, pediatric emergency medicine, and clinical pharmacy.
- The majority of the respondents agreed or strongly agreed that this measure do not have undue burden on hospital for its data. Respondents also noted that the data elements are currently available in a structured field in the electronic health record.
- There are no fees, licensing, or requirement for this measure.

Questions for the Committee:

o *Are the required data elements routinely generated and used during care delivery?*
o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
o *Is the data collection strategy ready to be put into operational use?*
o *If an eMeasure, does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility:

- This measure should not be computed using chart data given ADT data can track this information electronically. The focus on EHR or manual is still a barrier in measurement. However, the data is available in extraction.
- To my knowledge there is no significant burden to report and the required data elements are currently available in structured fields in the EHR.
- ED's that use EHR's should be able to collect this data easily. Facilities which use paper charts will face more challenged, but I do not think that is a significant barrier given the increasing prevalence of EHR's in ED's.
- Multiple Committee Members had no concerns.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- This measure is in use, publicly reported, and in use for benchmarking in Centers for Medicare and Medicaid Services Hospital Outpatient Quality Reporting program (HOQR) program.
- The data is publicly reported on the Hospital Compare Website as well as state and national averages are available for this measure. This data presents consumer the capability to compare the hospital's performance to other facilities and determine relative performance.
- The developer stated that the publicly reported values for all facilities must meet minimum case count requirements. Facilities that met the minimum case count requirements from January 2014-December 2016 ranged from 3,334 to 3,737 facilities per year. Facilities that report the requirement must follow the Outpatient Prospective Payment System guidelines.
- In the FY 2019 Inpatient Prospective Payment System (IPPS) proposed rule, CMS states:

  > With respect to the Median Time from ED Arrival to ED Departure for Discharged ED Patients measure (NQF 0496) (ED-3), this is an outpatient measure and is not included as an eCQM in the Hospital IQR Program. We are proposing to remove it so the eCQMs would align completely between the two programs in order to reduce burden and enable eligible hospitals and CAHs to easily report electronically through the Hospital IQR Program submission mechanism.
- CMS has not made any statements about removing the chart-abstracted version of this measure from the HOQR Program.
- The OPPS NPRM will be released in July, after which we will be able to provide more detail on future use of this measure. As of now, it remains in the HOQR Program.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others: N/A

Additional Feedback:

- Developer stated the reporting rate excludes psychiatric/mental health and transfer patients, which is not publicly reported on Hospital compare, but available at the website, https://data.medicare.gov/.

Questions for the Committee:

o *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*

o *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

- The developer indicated modest improvement in the median value time to discharged between January 2014 and December 2016 data periods. The median value time to discharge has declined by 2.9% or four minutes.
- The developer provided the following data on the improvement of facilities in the median value time to discharged:

	January 2014 – December 2014	**January 2016 – December 2016**
N	1,687,812	2,134,653
Met Minimum Case Count	3,334	3,737
Median Time to Discharge (minutes)	140	136

- The developer stated these cases reflect only a subset of patient eligible for this measure and were dependent on the facility's total case count; thus, the facility may report all cases or a sample of the cases.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- No unintended consequences were identified nor reported by external stakeholders since the implementation of this measure.
- Developer noted that potential unintended consequences will continue to be monitored through an annual review of the literature and an ongoing review of stakeholder comments and questions.

Potential harms

- No potential harms were identified for this measure.
- However, per the FY 2019 IPPS proposed rule, this measure could be removed to reduce hospitals' reporting burden.

Additional Feedback: N/A

Questions for the Committee:

o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

o *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency:

- Agree with pass rating. No concerns
- The measure is currently in the HOQR and an aspect of the measure, OP-18b, is used currently in the Star Ratings methodology as one measure under the Timeliness of Care group of measures.
- No major concerns here. How median ED time is used to change ED processes and how those changes impact clinical quality of care will be important going forward. The background material does not seem to address the impact of this measure on hospitals and patients.
- Publicly reported - not sure how otherwise used.
- Useful, though could be more useful for quality improvement if results were shared with EDs along strata where variations exist.
- The measure is publicly reported and hospitals are paid to report.

4b1. Usability – Improvement:

- No concerns
- Again, while use of this measure might help hospitals to work to improve ED throughput, comparisons without appropriate risk adjustment might frustrate large, urban hospitals with different ED utilization patterns and case-mix than the non-urban hospitals they are compared to. Performance results might be used to further the goal of high-quality, efficient care, but it simply might not be reasonable to expect large, urban hospitals to have the same ED throughput as smaller, non-urban facilities.
- Fundamentally, mean ED time is like a patient's temperature, blood pressure, or pulse: it means something, but what exactly it means and what should be done necessitates more complex discussion. There will certainly be unintended consequences of using this measure, but that should not prevent it's deployment. We track cycle time routinely in our family medicine clinics. What the "right" cycle time is depends on numerous variables. Median ED time is similar. The ACEP's experience with median time to antibiotic administration for pneumonia is a relevant example of how caution must be exercised with this type of measure.
- Don't think this is strong enough to be used as a performance measure.
- Useful
- Slight improvement has been seen from 2014 to 2016. There is a question on whether lower ED throughput time really reflects better care. Eliminating waiting is beneficial, but cutting into true evaluation time or treatment time could be deleterious. No unexpected results or unintended consequences were reported.

Criterion 5: Related and Competing Measures

Related or competing measures

- NQF #0495 : Median Time from ED Arrival to ED Departure for Admitted ED Patients
- NQF #0497: Admit Decision Time to ED Departure Time for Admitted Patients
- Left Without Being Seen is a CMS measure that calculates the percent of patients who leave the ED without being evaluated by a physician/advanced practice nurse/physician's assistant (physician/APN/PA).

Harmonization

- The developer stated that this measure is harmonized to the extent possible to the related and competing measures.
- Although the initial patient population are identified using different codes for NQF #0495 and NQF #0496, the difference is a function of the data availability rather than the clinical or methodological differences between these two measures.

- Meanwhile, NQF #0497 is an eCQM measure and reported the HIQR Program. However, its measure focus on the duration between the decision to admit a patient and the time the patient is discharged from the ED, which is a subset of a patient's total ED length of stay, as measured by NQF #0496.
- The target populations for NQF #0496 and the Left Without Being Seen are the same, but the Left Without Being Seen measure focuses on the percentage of patients that leave the ED without being seen by a physician/advanced practice nurse/physician's assistant (physician/APN/PA). The focus of the Left Without Being Seen measure differs from this measure.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

- None

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 19, 2018

- No comments have been submitted as of this date.

Measure Number: 0496
Measure Title: Median Time from ED Arrival to ED Departure for Discharged ED Patients

Evaluation A Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

   NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

   TIPS: *Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

   TIPS: *Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2

   TIPS: *Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

   I do have a concern that the developer should consider however. As noted in the previous committee evaluation, CMS is reporting the measure stratified by ED volume since that greatly influences the times and can skew the performance ratings. If this is indeed how the measure is being used in practice, it seems that testing of each stratified groups performance scores that will be reported is required? And I don't see any evidence that was tested.

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2

   TIPS: *Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☐High (go to Question #6)

   ☒Moderate (go to Question #6)

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.

   TIPS: *Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

TIPS: *For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

TIPS: *Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP: *You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:

☒High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)*

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

☐Low (please explain below) [NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required, but check with NQF staff]

*Based on testing that was done, note I am concerned not testing of measure rates stratified by volume as indicated is how it will be reported.

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

Under section 1.8 on testing for social risk factor adjustment, it indicates that 4 factors were tested and that "Results of the regression tests are reported in section **1b.4** of the Measure Submission Form." I don't see that anywhere? It later states under risk adjustment 2b3 section that no risk adjustment or risk stratification was performed. This is inconsistent? Where are the results? I do agree this measure probably should not be adjusted for social risk factors but still would like to see results. See also comments under Q13.

A second concern is around the inter-rater reliability cases testing comparing cases from chart abstraction to electronic data. In most case N = 12.410/2,343,102 = 0.5% or less than 1 percent of sample. Is this sufficient? It is concerning because all kappa/Pearson's correlation scores = 1.0. This is highly unusual; there is typically some small amount of disagreement due to simple coding or human error. Later on, they talk about "need to exclude any cases where the coder coded "UTD" or unable to determine to enable measure calculation." While this makes sense, my question is, does this mean there was something in the field in the actual data that was inconsistent with the medical record or was it missing in the data 100% of the time? If there was something in the field in the data and they excluded, they are artificially getting agreement by excluding the case when in fact there was disagreement. How are those cases to be excluded, if any exist, in real data? In this case, they were excluded based on the abstractors determination of UTD but that is not done in actual implementation. The number of cases where there was an abstract record were quite a small percentage of the sample as noted above, given the large amount of records for this measure. Given the small sample (0.5%), I am not sure if/how they are sure they got every type of potential error (also contributing to the odd 100% agreement).

____

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS*: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

Under section 1.8 on testing for social risk factor adjustment, it indicates that 4 factors were tested and that "Results of the regression tests are reported in section 1b.4 of the Measure Submission Form." I don't see that anywhere? It later states under risk adjustment 2b3 section that no risk adjustment or risk stratification was performed. This is inconsistent? Where are the results? I do agree this measure probably should not be adjusted for social risk factors but still would like to see results. I would like to see the results. It may point to the need to report rates stratified by some of those risk factors, like age or race/ethnicity, to give visibility to any disparities so that EDs can work to reduce them. I am not suggesting they be used to risk adjust the measures.

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with <u>patient-level data elements</u>?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: Prior validity studies of the same data elements may be submitted*

    ☒Yes (go to Question #22)

    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    REFERENCE: *Testing attachment, section 2b1.*

    TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

TIPS*: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Developer Responses:

The measure developers received a version of this measure worksheet in advance of committee review to check for factual accuracy and provide any clarifications. Regarding this review, the developer noted the following:

- In response to question #4: The current use of the measure in the HOQR Program does not stratify by ED volume. We note that the response on the prior Measure Testing Form, question 2b4.1 stated, “The results are stratified by reporting/non-reporting. The non-reporting group contains cases that were transferred or who had a psych diagnosis.” We provide more detailed information about the stratification in the 2018 MEF, section S.10. We confirm that testing was performed for the reported scores, and we included measure score reliability for all subgroups.
- In response to question #11: The UTD values are abstracted by both sites and the data validation contractor (i.e., there may have been errant or unclear data in the patient’s medical record, but both abstractors decided that these values were not usable for measure calculation and, thus, the case was removed from calculation of the facility’s score).

Measure Number: 0496

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

Evaluation B Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   TIPS*: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **<u>OR</u>** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with <u>computed performance measure scores</u> for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   TIPS*: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   Measure developers used data downloaded from hospital compare web site.

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2
   TIPS*: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   Demonstrated using the Intraclass Correlation (ICC) analysis with results displayed in exhibit 2 table.

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the <u>performance measure scores</u> are reliable?

   REFERENCE: Testing attachment, section 2a2.2
   TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☐High (go to Question #6)

   ☒Moderate (go to Question #6)

   Overall rate and reporting rate indicated high reliability while psychiatric and transfer cases indicated lower but acceptable ICC scores

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with <u>patient-level data elements</u> that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.
TIPS: *Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

Did not see evidence of admit date/time or discharge date/time data was tested for reliability. Measure developers appeared to have tested the results of the calculated time span but not the reliability of the actual data elements. I would recommend that the measure developers use the CDAC testing results to test reliability of date/time fields.

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2
TIPS: *For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

No explanation or results demonstrating reliability testing at the data element level.

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2
TIPS: *Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

No statistical analysis addressing missing data.

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS*: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☒Yes (please explain below then go to Question #13)

Measure developers did not discuss effect of UTD exclusions on validity. I would like the measure developers to consider replacing UTD cases from the overall qualifying population to assure that a minimal sample size is met.

☐No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

TIPS*: Consider the following:* ***If measure is risk adjusted****: If the developer asserts there is* ***no conceptual basis*** *for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model* *adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for* ***not risk adjusting*** *provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

I would like to see the measure developers analyze the impact of UTD data on the population sample. While the rates are low, I do not believe that is enough to conclude that there is no effect on the overall results. In addition, it would be helpful for the developers to add a section describing how the sample size is determined.

☐No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☒Yes (please explain below then go to Question #17)

Yes, see previous comments to UTD cases

☐No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☐Yes (go to Question #20)

☒No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

While the validity of the performance score demonstrates a difference, you cannot conclude that falling above or below the median is better or worse performance. One could conclude that falling below the median means the provider did not spend enough time with the patient which would indicate poor performance. In contrast, the same patient falling above the median could indicate that the provider spent more quality time with the patient. There is no way to demonstrate good or bad performance, only that there are differences in the median times. I would recommend the developers do two things; 1) Consider adding risk adjustment i.e. comorbidities, dual eligibility and 2) Add supporting references supporting how length of time in the emergency department impacts patient outcomes. This would at least support that using this measure as a proxy measure for quality of care is supported in current literature.

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☐Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☒Insufficient (go to Question #21)

21. Was validity testing conducted with <u>patient-level data elements</u>?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements?

NOTE that data element validation from the literature is acceptable.

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS: *Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS: *Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

While I believe the measure is valid, I also believe references to good and bad performance needs to be clearly defined with strong, supporting evidence. Otherwise, we are assuming a shorter stay in the ED is the better outcome.

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Developer Responses:

The measure developers received a version of this measure worksheet in advance of committee review to check for factual accuracy and provide any clarifications. Regarding this review, the developer noted the following:

- In response to question #33: For reliability testing, we tested the performance score ("calculated time span"). The reliability of the actual data elements is addressed by the data element validity testing included in section 2b3.1, in accordance with the MTF instructions. We tested data element validity of arrival time, ED departure date, and ED departure by comparing the CDW data to CDAC, as recommended by the reviewer.
- In response to question #38: Because the measure is chart abstracted, we do not receive cases for which values were missing (they are rejected before inclusion in the measure population); if there are value(s) for which abstractors are unable to provide data, they must select *UTD* to remove a case. Because of this structure, we are unable to evaluate the difference in cases included vs. those that are not (as abstraction ends once a case hits *UTD* for any data element).
- In response to question #39: This is a byproduct of the data that we receive for chart-abstracted cases. In order to assess what the methods panel reviewer suggests, we would need to ask abstractors to collect data for all elements (even after selecting *UTD*), which would be unduly burdensome. A description of sample size determinations is provided in MSF, section S.15.
- In response to question #46: Evidence supporting decreased throughput times improves patient outcomes is provided for the *Evidence* criterion and that our expert work group did not recommend any additional factors by which the measure should be stratified (beyond patients seeking mental/behavioral health services and transfers).

Measure Number: 0496

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

Evaluation C Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   TIPS*: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **<u>OR</u>** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with <u>computed performance measure scores</u> for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   TIPS*: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2
   TIPS*: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the <u>performance measure scores</u> are reliable?

   REFERENCE: Testing attachment, section 2a2.2
   TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☐High (go to Question #6)

   ☒Moderate (go to Question #6)

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with <u>patient-level data elements</u> that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.
   TIPS*: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

TIPS: *For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

TIPS: *Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS: *Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS: *Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☒No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

TIPS: *Consider the following:* ***If measure is risk adjusted****: If the developer asserts there is* **no conceptual basis** *for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for* **not risk adjusting** *provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

The provided materials discuss prior discussion of whether some adjustment should be made for differences in case mix across EDs that might legitimately change the appropriate time needed to assess and treat. There is no evidence of assessment of whether case mix adjustment would change scores or rankings. This should be part of the substantive committee's review.

The average ICC for the psychiatric subgroup was acceptable but some samples generated notably lower ICC's. The psych strata is a more fragile strata, although I'm not prepared to reject it.

☐No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

TIPS*: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20) No comparisons are presented, but this is a process measure and meets the face validity and data validity checks needed to assure it is measuring ED median time. Whether ED median time is a useful measure I leave to the substantive committee.

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

TIPS: *Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements?

NOTE that data element validation from the literature is acceptable.

REFERENCE: *Testing attachment, section 2b1.*

TIPS: *For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?
REFERENCE: Testing attachment, section 2c
***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0496
Corresponding Measures:
De.2. **Measure Title:** Median Time from ED Arrival to ED Departure for Discharged ED Patients
Co.1.1. **Measure Steward:** Centers for Medicare and Medicaid Services
De.3. **Brief Description of Measure:** NQF #0496 calculates the median time from emergency department arrival to time of departure from the emergency room for patients discharged from the emergency department (ED). The measure is calculated using chart-abstracted data, on a rolling quarterly basis, and is publically reported in aggregate for one calendar year. The measure has been publically reported since 2013 as part of the ED Throughput measure set of the CMS' Hospital Outpatient Quality Reporting (HOQR) Program.
1b.1. **Developer Rationale:** Empirical evidence demonstrates that ED throughput is an indicator of hospital quality of care, and shows that shorter lengths of stay in the ED lead to improved clinical outcomes. Significant ED overcrowding has numerous downstream effects, including prolonged patient waiting times, increased suffering for those who wait, rushed and unpleasant treatment environments, and potentially poor patient outcomes (Gardner, 2018). Quality improvement efforts aimed at reducing ED overcrowding and length of stay have been associated with an increase in ED patient volume, decrease in number of patients who leave without being seen, reduction in costs, and increase in patient satisfaction (Bucci, 2016; Chang, 2017; Zocchi, 2015). An analysis of data from 2,619 hospitals support that reducing the time patients remain in the ED is associated with increased patient satisfaction and a decreased chance that patients will leave before being seen (Chang, 2017). Recent guidelines and peer-reviewed studies also demonstrate the need for dedicated emergency mental health services, providing evidence that the clinical needs for these patients substantively differ from the non-psychiatric population (Nazarian, 2017; Lester, 2018).

REFERENCES:
1) Bucci, S., A. G. de Belvis, S. Marventano, A. C. De Leva, M. Tanzariello, M. L. Specchia, W. Ricciardi and F. Franceschi. (2016). Emergency department crowding and hospital bed shortage: Is Lean a smart answer? A systematic review. Eur Rev Med Pharmacol Sci, 20(20), 4209-4219.
2) Chang, A. M., A. Lin, R. Fu, K. J. McConnell and B. Sun. (2017). Associations of Emergency Department Length of Stay With Publicly Reported Quality-of-care Measures. Acad Emerg Med, 24(2), 246-250.
3) Gardner, R. M., N. A. Friedman, M. Carlson, T. S. Bradham and T. W. Barrett. (2017). Impact of revised triage to improve throughput in an ED with limited traditional fast track population. Am J Emerg Med., 36(1), 124-127.
4) Lester, N. A., L. R. Thompson, K. Herget, J. A. Stephens, J. V. Campo, E. J. Adkins, T. E. Terndrup and S. Moffatt-Bruce. (2017). CALM Interventions: Behavioral Health Crisis Assessment, Linkage, and Management Improve Patient Care. Am J Med Qual., 33(1), 65-71.
5) Nazarian DJ, Broder JS, Thiessen ME, Wilson MP, Zun LS, Brown MD, American College of Emergency Physicians. Clinical policy: critical issues in the diagnosis and management of the adult psychiatric patients in the emergency department. Ann Emerg Med. 2017 Apr; 69(4):480-98. Guideline available at: http://www.annemergmed.com/article/S0196-0644(17)30070-7/pdf.
6) Zocchi, M. S., M. S. McClelland, and J. M. Pines. Increasing Throughput: Results From A 42-Hospital Collaborative To Improve Emergency Department Flow. The Joint Commission Journal on Quality and Patient Safety, 2015, 41(12):532–542.

S.4. **Numerator Statement:** Continuous Variable Statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.
S.6. **Denominator Statement:** This measure is reported as a continuous variable statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.
S.8. **Denominator Exclusions:** Patients who expired in the emergency department, left against medical advice (AMA), or whose discharge was not documented or unable to be determined (UTD) are excluded from the target population.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
S.20. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Sep 09, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable; this measure is not a paired or grouped measure.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_0496_Measure_Evidence_Form.docx

1a.1 <u>For Maintenance of Endorsement:</u> Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Empirical evidence demonstrates that ED throughput is an indicator of hospital quality of care, and shows that shorter lengths of stay in the ED lead to improved clinical outcomes. Significant ED overcrowding has numerous downstream effects, including prolonged patient waiting times, increased suffering for those who wait, rushed and unpleasant treatment environments, and potentially poor patient outcomes (Gardner, 2018). Quality improvement efforts aimed at reducing ED overcrowding and length of stay have been associated with an increase in ED patient volume, decrease in number of patients who leave without being seen, reduction in costs, and increase in patient satisfaction (Bucci, 2016; Chang, 2017; Zocchi, 2015). An analysis of data from 2,619 hospitals support that reducing the time patients remain in the ED is associated with increased patient satisfaction and a decreased chance that patients will leave before being seen (Chang, 2017). Recent guidelines and peer-reviewed studies also demonstrate the need for dedicated emergency mental health services, providing evidence that the clinical needs for these patients substantively differ from the non-psychiatric population (Nazarian, 2017; Lester, 2018).

REFERENCES:

1) Bucci, S., A. G. de Belvis, S. Marventano, A. C. De Leva, M. Tanzariello, M. L. Specchia, W. Ricciardi and F. Franceschi. (2016). Emergency department crowding and hospital bed shortage: Is Lean a smart answer? A systematic review. Eur Rev Med Pharmacol Sci, 20(20), 4209-4219.

2) Chang, A. M., A. Lin, R. Fu, K. J. McConnell and B. Sun. (2017). Associations of Emergency Department Length of Stay With Publicly Reported Quality-of-care Measures. Acad Emerg Med, 24(2), 246-250.

3) Gardner, R. M., N. A. Friedman, M. Carlson, T. S. Bradham and T. W. Barrett. (2017). Impact of revised triage to improve throughput in an ED with limited traditional fast track population. Am J Emerg Med., 36(1), 124-127.

4) Lester, N. A., L. R. Thompson, K. Herget, J. A. Stephens, J. V. Campo, E. J. Adkins, T. E. Terndrup and S. Moffatt-Bruce. (2017). CALM Interventions: Behavioral Health Crisis Assessment, Linkage, and Management Improve Patient Care. Am J Med Qual., 33(1), 65-71.

5) Nazarian DJ, Broder JS, Thiessen ME, Wilson MP, Zun LS, Brown MD, American College of Emergency Physicians. Clinical policy: critical issues in the diagnosis and management of the adult psychiatric patients in the emergency department. Ann Emerg Med. 2017 Apr; 69(4):480-98. Guideline available at: http://www.annemergmed.com/article/S0196-0644(17)30070-7/pdf.

6) Zocchi, M. S., M. S. McClelland, and J. M. Pines. Increasing Throughput: Results From A 42-Hospital Collaborative To Improve Emergency Department Flow. The Joint Commission Journal on Quality and Patient Safety, 2015, 41(12):532–542.

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Analysis of facility-level data from the Hospital Compare downloadable files indicates that there is variation in the median time from ED arrival to time of departure from the emergency room. During the January 2014 to December 2014 data collection periods, median facility-level throughput times ranged from 46 minutes to 424 minutes, with a median of 140 minutes. During the January 2016 to December 2016 data collection periods, median facility-level throughput times ranged from 45 minutes to 440 minutes, with a median of 136 minutes. When aggregating findings from both data collection periods, the median value of median time from emergency department arrival to time of departure from the emergency room decreased 2.9% (-4 minutes).

The data presented below represent performance scores and descriptive statistics for longitudinal facility performance for the facilities whose denominator counts met minimum case count requirements during the January 2014 to December 2016 data collection periods.

	Data Collection Period			Change in Minutes
	January 2014–December 2014	January 2015–December 2015	January 2016–December 2016	2014–2016
Facilities	3,334	3,584	3,737	-
Minimum Value	46	49	45	-1
1st Percentile	74	70	68	-6
5th Percentile	88	84	84	-4
10th Percentile	98	94	94	-4
25th Percentile	116	114	112	-4
Median	140	138	136	-4
75th Percentile	167	166	165	-2
90th Percentile	195	196	196	+1
95th Percentile	218	218	217	-1
99th Percentile	272	264	266	-6
Maximum Value	424	428	440	+16
Number of ED cases (Denominator)	1,687,812	1,870,875	2,134,653	-

During the January 2014 to December 2016 data collection periods, there is documentation of substantial variation in facility performance. The interquartile range has been consistently wide, ranging from 112 minutes to 165 minutes. Additionally, the maximum time for ED discharge increased between 2014 and 2016. While median performance is improving, there is an ongoing opportunity for improvement in performance at the facility level.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Data have been included in Section 1b.2; these data represent national performance over time, from the January 2014 to December 2016 data collection periods.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(<u>This is required for maintenance of endorsement</u>. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
The relationship of patient and facility characteristics on ED throughput time was evaluated using an Ordinary Least Squares (OLS) regression (with standard errors clustered at the facility level), using data submitted to the Clinical Data Warehouse (CDW) between October 01, 2015 and August 30, 2016. Appendix A describes the methods and regression results for the Overall Rate and the three strata. Results from the Reporting Rate regression are summarized below. It is important to note that, while many

results are significant, this may be driven by sample size at the facility level; the magnitude of these differences may not be clinically meaningful.

Primary results from the regression were related to patient demographics. ED throughput time was significantly longer for patients in the 18–30 (ß= 31.4 minutes, p<0.001), 30–40 (ß= 41.1 minutes, p<0.001), 40–50 (ß= 53.2 minutes, p<0.001), 60–70 (ß= 70.1 minutes, p<0.001), 70–80 (ß= 77.3 minutes, p<0.001), 80–90 (ß= 84.9 minutes, p<0.001), and over 90 (ß= 91.6 minutes, p<0.001) age groups, as compared to those patients less than 18 years old. There was a significantly longer ED throughput times for female patients, as compared to male patients (ß= 6.1 minutes, p< 0.001). When compared to white patients, there was a significantly longer ED throughput time for Asian patients (ß= 10.3 minutes, p< 0.001); Hispanic patients also experienced longer ED throughput times, as compared to the non-Hispanic peers (ß= 12.0 minutes, p<0.001).

ED throughput times also varied by the characteristics of the facility from which the patient was discharged. When compared to patients discharged from facilities with fewer than 50 beds (a proxy for facility size), there was a significantly longer ED throughput time for patients discharged from facilities with 51–100 beds (ß= 10.7 minutes, p=0.004), 101–250 beds (ß= 35.8 minutes, p <0.001), 251–500 beds (ß= 53.8 minutes, p <0.001), and more than 500 beds (ß= 64.7 minutes, p< 0.001). Urbanicity also impacted ED throughput times, with a significantly higher time for patients discharged from an urban hospital, as compared to those discharged from a rural hospital (ß= 6.5 minutes, p< 0.001). Finally, when compared to patients discharged from a non-teaching facility, there was a significantly longer ED throughput time for patients discharged from a major teaching facility (ß= 54.7 minutes, p< 0.001).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**

The Centers for Medicare and Medicaid Services (CMS) released findings showing variation in hospital performance for several ED throughput metrics, including wait time (Sun et al. 2016). The findings were based on a risk-adjusted model that included variables related to hospitals' structural, financial, and geographic characteristics. Risk adjustment is the statistical process used to adjust for differences in population or setting characteristics before comparing outcomes. The statistical model used by Sun et al. (2016) included several hospital characteristics; their doing so acknowledged and controlled for the impact of hospital-level characteristics on patient outcomes. After risk adjusting the measures based on hospital characteristics, variations in ED throughput existed; these findings support the need for ongoing reporting of the ED throughput measures, including NQF #0496.

REFERENCES:
1) Sun, B. C., A. Laurie, L. Prewitt, R. Fu, A. M. Chang, J. Augustine, C. t. Reese and K. J. McConnell (2016). "Risk-Adjusted Variation of Publicly Reported Emergency Department Timeliness Measures." Annals of Emergency Medicine, 67(4), 509-516 e7.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Non-Condition Specific***(check all the areas that apply):*
Care Coordination

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, Elderly, Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Women

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FSpecsManualTemplate&cid=1228776146046

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_0496_Measure_Code_Set.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
Yes

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
NQF #0496 was first endorsed by NQF in October 2008. Since 2008, its measure specifications have been updated to address stakeholder feedback and to harmonize with measures in the Hospital Inpatient Quality Reporting (HIQR) program. Some data elements have been updated to provide clarification in abstraction and updates have been made to selected references since the measure's most recent NQF review, which occurred in 2014.
In 2015, as part of the annual measure maintenance and review process, all ICD-9-CM diagnosis codes were updated to align with corresponding ICD-10-CM diagnosis codes. For all subsequent years, the list of ICD-10-CM diagnosis codes used to identify psychiatric/mental health rate cases have been updated annually to align with CMS ICD-10 diagnosis codes and descriptions.
In 2016, the Arrival Time data element was modified to specify that the ED record may include the ED face sheet, ED consent or authorization for treatment forms, ED or outpatient registration or sign-in forms, ED ECG reports, ED telemetry or rhythm strips, ED laboratory reports, and ED X-ray reports. In 2017, guidance was added to the ED Departure Time data element to clarify the rationale for abstracting the observation order. Guidance was also added to the Discharge Code data element to clarify when to abstract value a value of 7—Left Against Medical Advice.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Continuous Variable Statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The measure population is identified using six evaluation and management (E/M) codes for ED encounters. ICD-10-CM diagnosis codes and discharge codes are used to identify cases for the Psychiatric/Mental Health Rate and Transfer Rate strata. These detailed lists can be found in the Excel workbook provided for Section S.2b.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
This measure is reported as a continuous variable statement: Time (in minutes) from ED arrival to ED departure for patients discharged from the emergency department.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
NQF #0496 is a continuous measure; therefore, the numerator and denominator details contained in Section S.6 and Section S.9 are the same.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients who expired in the emergency department, left against medical advice (AMA), or whose discharge was not documented or unable to be determined (UTD) are excluded from the target population.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
The Discharge Code data element is used to identify measure exclusions [Discharge Code equals: 6—Expired, 7—Left Against Medical Advice/AMA, or 8—Not Documented or Unable to Determine (UTD)].

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
NQF #0496 is specified using an overall rate, with three sub-populations (or strata), described in detail in Section 1.2 of the Measure Testing Form, and summarized 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, provided in Attachment: NQF_0496_Measure Code Set.xlsx.
- 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.

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 sociodemographic factors; doing so would potentially mask 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.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Continuous variable
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
This measure calculates the time (in minutes) from ED arrival to ED departure for discharged ED patients. The patient population is determined from two algorithms: the Hospital Outpatient ED Throughput Population algorithm as well as the NQF #0496 measure-specific algorithm:

1. Start processing. Run all cases that are included in the ED Throughput Hospital Outpatient Population Algorithm and pass the edits defined in the Data Processing Flow through this measure. Proceed to ICD-10-CM Principal Diagnosis Code.

2. Check Discharge Code.

a. If Discharge Code is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If Discharge Code equals 6, 7, or 8 the case will proceed to a Measure Category Assignment of B. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

c. If Discharge Code equals 1, 2, 3, 4a, 4b, 4c, 4d, or 5, the case will proceed to Arrival Time.

3. Check Arrival Time.

a. If Arrival Time equals UTD, the case will proceed to a Measure Category Assignment of Y. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If Arrival Time equals non-UTD value, the case will proceed to ED Departure Date.

4. Check ED Departure Date.

a. If ED Departure Date is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If ED Departure Date equals UTD, the case will proceed to a Measure Category Assignment of Y. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

c. If ED Departure Date equals non-UTD, the case will proceed to ED Departure Time.

5. Check ED Departure Time.

a. If ED Departure Time is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If ED Departure Time equals UTD, the case will proceed to a Measure Category Assignment of Y. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

c. If ED Departure Time equals non-UTD, the case will proceed to Measurement Value.

6. Calculate the Measurement Value. Time in minutes is equal to the ED Departure Date and ED Departure Time (in minutes) minus the Outpatient Encounter Date and Arrival Time (in minutes).

7. Check Measurement Value.

a. If Measurement Value is less than 0 minutes, the case will proceed to a Measure Category Assignment of X and will be rejected. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If Measurement Value is greater than or equal to 0 minutes, the case will proceed to a Measure Category Assignment of D1.

8. Initialize the Measure Category Assignment for all cases in D1.

9. Proceed to ICD-10-CM Principal Diagnosis Code.

10. Check ICD-10-CM Principal Diagnosis Code.

a. If ICD-10-CM Principal Diagnosis Code is in Appendix A, OP Table 7.01 of the HOQR Specifications Manual (refer to Attachment: NQF_0496_Measure Code Set.xlsx for corresponding ICD-10 codes), the case will proceed to a Measure Category Assignment of D2. Proceed to Discharge Code.

b. If ICD-10-CM Principal Diagnosis Code is not in Appendix A, OP Table 7.01, the case will proceed to Discharge Code.

11. Check Discharge Code.

a. If Discharge Code equals 4a or 4d, the case will proceed to a Measure Category Assignment of D3. Proceed to ICD-10-CM Principal Diagnosis Code.

b. If Discharge Code equals 1, 2, 3, 4b, 4c, or 5, the case will proceed to ICD-10-CM Principal Diagnosis Code.

12. Check ICD-10-CM Principal Diagnosis Code.

a. If ICD-10-CM Principal Diagnosis Code is in Appendix A, OP Table 7.01, the case will proceed to a Measure Category Assignment of B. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

b. If ICD-10-CM Principal Diagnosis Code is not in Appendix A, OP Table 7.01, the case will proceed to Discharge Code.

13. Check Discharge Code.

a. If Discharge Code equals 4a or 4d the case will proceed to a Measure Category Assignment of B. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.
If Discharge Code equals 1, 2, 3, 4b, 4c, or 5, the case will proceed to a Measure Category Assignment of D. Return to Transmission Data Processing Flow: Clinical in the Data Transmission section.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Sampling is a process of selecting a representative subset of a population in order to estimate the hospital's overall performance without collecting data for its entire population. Using a statistically valid sample, a hospital can measure its performance in an effective and efficient manner. Sampling is a particularly useful technique for performance measures that require primary data collection from a source such as the medical record. In order to avoid having outliers exert a disproportionate effect on the estimate of overall performance, sampling should not be used unless the hospital has a sufficiently large number of cases in the universe of interest. For the purpose of sampling outpatient department quality measures, the terms "sample," "effective sample," and "case" are defined below:

- The "sample" is the fraction of the population that is selected for further study.
- "Effective sample" refers to the part of the sample remaining after application of exclusion and exception criteria. It is defined as the sample for an outpatient measure set minus all the exclusions and contraindications for the outpatient measure set in the sample. The effective sample serves as the denominator population of an outpatient measure.
- A "case" refers to a single record (or an encounter) within the population. For example, during the first quarter a hospital may have 100 patients who had a principal diagnosis associated with the ED Throughput measures. The hospital's outpatient population would include 100 cases or 100 outpatient records for these measures during the first quarter.

To obtain statistically valid sample data, the sample size should be carefully determined, and the sample cases should be randomly selected in such a way that the individual cases in the population have an equal chance of being selected. Only when the sample data truly represent the whole population can the sample-based performance outpatient measure data be meaningful and useful. Each hospital is ultimately responsible for adhering to the sampling requirements outlined in Section 4 of the Hospital OQR Specifications Manual, to which we have linked in Section S.1.

As a general rule/policy of CMS, providers are encouraged to exceed minimum sampling requirements and submit as many cases as possible up to the entire population of cases if reasonably feasible. For example, if the raw data can be easily extracted from an existing electronic database or the abstraction burden is manageable, providers should consider submitting the entire population of cases that meet the initial selection criteria. Otherwise, a statistically valid sample can be selected for reporting.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
This measure does not use survey or patient-reported data.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
An electronic data collection tool, CMS Abstraction & Reporting Tool (CART), is available for third-party vendors or facilities to download for free. Paper tools for manual abstraction, which are posted on www.qualitynet.org, are also available for the CART tool.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Emergency Department and Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable; this is not a composite measure.

2. Validity – See attached Measure Testing Submission Form
NQF_0496_Measure_Testing_Form.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
Not applicable; the data elements are in defined fields in a combination of electronic sources. The potential for electronic specification will require special attention to the Arrival Time, ED Departure Date, and ED Departure Time data elements. Abstractors rely on documentation in the ED record to determine the time the patient physically arrived in and left the ED.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Nine expert work group (EWG) members, with backgrounds in healthcare administration, management, and clinical expertise in emergency medicine, pediatric emergency medicine, and clinical pharmacy, provided feedback on the feasibility of NQF #0496 through an online survey. Most respondents agreed or strongly agreed that the practical aspects of reporting NQF #0496 chart-abstracted measure do not place undue burden on hospitals for its data. However, one respondent commented that the degree of burden may vary depending on the programming structure of different electronic health records (EHRs). Most respondents also indicated that the data elements are currently available in an electronic health record EHR structured field. Overall, the respondents generally support the feasibility of NQF #0496.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm).*
No fees, licensure, or other requirements are necessary to use this measure; however, CPT codes, descriptions, and other data are copyright 2013 American Medical Association. All rights reserved. CPT is a registered trademark of the American Medical Association. Applicable FARS\DFARS Restrictions Apply to Government Use. Fee schedules, relative value units, conversion factors, and/or related components are not assigned by the AMA, are not part of CPT, and the AMA is not recommending their use. The AMA does not directly or indirectly practice medicine or dispense medical services. The AMA assumes no liability for data contained or not contained herein.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)

Public Reporting Quality Improvement (Internal to the specific organization)	Payment Program CMS HIQR Program https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier2&cid=1138115987129 Regulatory and Accreditation Programs Joint Commission Accreditation http://www.jointcommission.org/accreditation_process_overview/

4a1.1 **For each CURRENT use, checked above (update for maintenance of endorsement), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

Public Reporting
Name of program and sponsor: CMS HOQR Program
Purpose: The HOQR Program is a pay for quality data reporting program implemented by CMS for outpatient hospital services. Hospital quality of care information gathered through the HOQR Program is publicly available on the Hospital Compare website. In addition to providing hospitals with a financial incentive to report their quality of care measure data, the HOQR Program provides CMS with data to help Medicare beneficiaries make more informed decisions about their health care.
Geographic area and number and percentage of accountable entities and patients included: The publicly reported values (on Hospital Compare) are calculated for all facilities in the United States that meet minimum case count requirements. The number of facilities that met minimum case count criteria during the January 2014 to December 2016 data collection periods ranged from 3,334 to 3,737 facilities annually. The number of facilities meeting minimum case count criteria by year is presented in Section 1b.2. Facilities eligible to report this measure are subject to the Outpatient Prospective Payment System (OPPS) guidelines.
Level of measurement and setting: Facility; Emergency Department and Services

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
Name of program and sponsor: CMS HOQR Program
Purpose: The HOQR Program is a pay-for-quality data reporting program implemented by CMS for outpatient hospital services. In addition to providing hospitals with a financial incentive to report their quality of care measure data, the data is publicly reported on the Hospital Compare Website. The data reported on Hospital Compare not only shows the hospital's score on the measure, but also provides state and national averages for the measure. This enables consumers to compare the hospital's performance to other facilities and determine relative performance.
Geographic area and number and percentage of accountable entities and patients included: The publicly reported values (on Hospital Compare) are calculated for all facilities in the United States that meet minimum case count requirements. The number of facilities that met minimum case count criteria during the January 2014 to December 2016 data collection periods ranged from 3,334 to 3,737 facilities, annually. The number of facilities meeting minimum case count criteria by year is presented in Section 1b.2.Facilities eligible to report this measure are subject to the OPPS guidelines.
Level of measurement and setting: Facility; Emergency Department and Services

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The reporting rate, which excludes psychiatric/mental health and transfer patients, is publicly reported on Hospital Compare. Although the Psychiatric/Mental Health Rate will not be publicly reported on Hospital Compare, it will be publically available at https://data.medicare.gov/.

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
This measure is publicly reported.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Not applicable

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
Not applicable

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Not applicable

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Not applicable

4a2.2.3. **Summarize the feedback obtained from other users**
Not applicable

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
Not applicable

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
Summary statistics of performance scores during the January 2014 to December 2016 data collection periods are provided in Section 1b.2. The median value time to discharge has declined 2.9% (4 minutes) between the January 2014 and December 2016 data collection periods. There were 3,334 facilities that met minimum case count during the January 2014 to December 2014 data collection periods; 3,737 facilities met the minimum case count for the January 2016 to December 2016 data collection periods. During the January 2014 to December 2014 data collection periods, there were 1,687,812 sampled cases; of those patients, the median time to discharge was 140 minutes. During the January 2016 to December 2016 data collection periods, there were 2,134,653 sampled cases; of those patients, the median time to discharge was 136 minutes. These cases reflect only a subset of the patients eligible for the measure. Depending on the facility's total case count, the facility may report all cases or a sample of cases.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
Measure testing did not identify any unintended consequences. Similarly, no evidence of unintended consequences to individuals or populations has been reported by external stakeholders since its implementation. The potential for unintended consequences will continue to be monitored through an annual review of the literature as well as an ongoing review of stakeholder comments and inquiries.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
No unexpected benefits have been identified after implementing the measure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0495 : Median Time from ED Arrival to ED Departure for Admitted ED Patients

0497 : Admit Decision Time to ED Departure Time for Admitted Patients

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

Left Without Being Seen is a CMS measure that calculates the percent of patients who leave the ED without being evaluated by a physician/advanced practice nurse/physician's assistant (physician/APN/PA).

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

The measure specifications are harmonized to the extent possible; however, the differences are justified. NQF #0496 is reported through the HOQR Program as a chart-abstracted measure, while NQF #0495 (Median Tine from ED Arrival to ED Departure for Admitted Patients) is reported through the Hospital Inpatient Quality Reporting (HIQR) Program as an electronically specified clinical quality measure (eCQM). Although the initial patient populations are identified using different codes, the difference is a function of data availability rather than clinical or methodologic differences in the populations measured by NQF #0496 and NQF #0495. NQF #0497 (Median Admit Decision Time to ED Departure Time for Admitted Patients) is also an eCQM, reported through the HIQR Program. Its measure focus is the duration between the decision to admit a patient and the time the patient is discharged from the ED, which is a subset of a patient's total ED length of stay, as measured by NQF #0496. While the target populations for NQF #0496 and Left Without Being Seen are the same, the focus of the measures is different. NQF #0496 focuses on the median time from ED arrival to ED departure for discharged patients, while Left Without Being Seen focuses on the percentage of patients that leave the ED without being seen by a physician/advanced practice nurse/physician's assistant (physician/ APN/ PA).

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

No competing measure that address both the same measure focus and target population as NQF #0496 was identified.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** NQF_0496_Appendix.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare and Medicaid Services
Co.2 **Point of Contact:** Joseph, Clift, joseph.clift@cms.hhs.gov, 410-786-4165-
Co.3 **Measure Developer if different from Measure Steward:** The Lewin Group
Co.4 **Point of Contact:** Colleen, McKiernan, Colleen.McKiernan@lewin.com, 703-269-5595-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The contractor has convened an EWG, which evaluates and provides feedback on measure-development and maintenance efforts for throughput measures. Specifically, the EWG provides direction and feedback through all phases of project activities, including expansion of the measures to additional CMS quality reporting programs, updates to the current specifications of these four measures, review of quantitative testing results, feedback on qualitative testing questions (i.e., results of EWG member questionnaires), and support for endorsement of the measures by NQF.

The following is a list of the contractor's EWG members:

- Anthony Arguija, MD, Long Beach Memorial Medical Center, Department of Emergency Medicine
- Bradley Weiner, MD, American Orthopaedic Association (AOA)
- Cathy Olson, MSN, RN, Emergency Nurses Association (ENA), Institute for Quality, Safety, and Injury Prevention, Director
- Daniel Waxman, MD, RAND Corporation
- David Marcozzi, MD, MHS-CL, FACEP, University of Maryland School of Medicine
- David Ring, MD, PhD, American Academy of Orthopaedic Surgeons (AAOS)
- Jeffrey A. Seiden, MD, Children´s Hospital of Philadelphia (CHOP)
- John Couk, MD, Louisiana State University Health Care Services Division (HCSD)
- Mary Ann Kliethermes, Pharm D, American Pharmacists Association (APhA)
- Matt Zavadsky, MS-HPA, National Association of Emergency Medical Technicians (NAEMT)
- Richard Newell, MD MPH FACEP, CEP America, Director of Quality and Performance
- Stephen Traub, MD, TEP 2010; Mayo Clinic, Department of Emergency Medicine, Chair
- Bradford Tinloy MD, CEP America, Medical Director of CMS Programs

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2008
Ad.3 **Month and Year of most recent revision:** 09, 2017
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 09, 2018

Ad.6 **Copyright statement:** This measure does not have a copyright.
Ad.7 **Disclaimers:** CPT codes, descriptions, and other data only are copyright 2013 American Medical Association. All rights reserved. CPT is a registered trademark of the American Medical Association. Applicable FARS\DFARS Restrictions Apply to Government Use. Fee schedules, relative value units, conversion factors and/or related components are not assigned by the AMA, are not part of CPT, and the AMA is not recommending their use. The AMA does not directly or indirectly practice medicine or dispense medical services. The AMA assumes no liability for data contained or not contained herein.

Ad.8 **Additional Information/Comments:**

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

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

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 4/16/2018

Please note: 2014 submission text in black | 2018 submission text in red

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ ☒ Process: This measure calculates the median time from emergency department (ED) arrival to time of departure from the emergency room for patients discharged from the emergency department.

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

2014 Submission: Not Applicable

2018 Submission: NQF #0496 measures the median time from ED arrival to ED departure, which documents a patient's length of stay in the emergency department. Facilities that report a high median time from arrival to departure may experience significant ED crowding, which is associated with unfavorable health outcomes, including longer hospital stays, increased costs, and higher mortality rates (Sun et al., 2013). By improving ED throughput times, facilities can increase ED patient volume, decrease the number of patients who leave without being seen, reduce costs, and increase patient satisfaction (Bucci et al., 2016; Chang et al., 2017; Zocchi et al., 2015).

REFERENCES:

1) Bucci, S., A. G. de Belvis, S. Marventano, A. C. De Leva, M. Tanzariello, M. L. Specchia, W. Ricciardi and F. Franceschi. Emergency department crowding and hospital bed shortage: Is Lean a smart answer? A systematic review. Eur Rev Med Pharmacol Sci, 2016, 20(20), 4209-4219.
2) Chang, A. M., A. Lin, R. Fu, K. J. McConnell and B. Sun. (2017). Associations of Emergency Department Length of Stay With Publicly Reported Quality-of-care Measures. Acad Emerg Med, 24(2), 246-250.
3) Zocchi, M. S., M. S. McClelland, and J. M. Pines. Increasing Throughput: Results From A 42-Hospital Collaborative To Improve Emergency Department Flow. The Joint Commission Journal on Quality and Patient Safety, 2015, 41(12):532–542.

4) Sun, B.C., Hsia, RY, Weis, RE, Zingmond, D, Liang, L.J., Han, W., McCreath, H., Asch, S.M. Effect of emergency department crowing on outcomes of admitted patients. Annals of Emergency Medicine, 2013 Jun, 61(6):605-611.

1a.3 Value and Meaningfulness: **IF** this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

2014 Submission: *Blank – new question*

2018 Submission: Not applicable

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number**	

- **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: - Quantity – how many studies? - Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 **Briefly SYNTHESIZE the evidence that supports the measure.** A list of references without a summary is not acceptable.

2014 Submission: Not Applicable
2018 Submission: The measure developer conducts an annual review of clinical practice guidelines and the peer-reviewed literature to identify evidence and/or new studies that relate to the measure or its clinical intent. Citations and summaries for seven reports included in this review can be found in **Section 1a.4.3.**

The evidence base for NQF #0496 shows that ED throughput is a meaningful indicator of hospital quality of care, and validates that shorter ED lengths of stay lead to improved clinical outcomes (Gardner et al. 2018). Mullins et al. studied data from *Hospital Compare*, which use the *Reporting Rate* strata for NQF #0496; the research team concluded that there is widespread variation in performance across the United States and that ED crowding is linked to inpatient quality outcomes (2014). An analysis of data from 2,619 hospitals showed that reducing ED length of stay is associated with increased patient satisfaction and decreased likelihood that a patient will leave before a medical professional sees him or her (Chang et al. 2017). Authors of multiple studies describe quality improvement and Lean-based interventions, which aim to improve ED throughput time and show that ED crowding and timely throughput remain high-priority issues for hospitals (Melton et al. 2016; Allaudeen et al. 2017; Bucci et al. 2016). A 2017 guideline prepared by the American College of Emergency Physicians (ACEP) justifies the separate measurement of patients for mental health and psychiatric services (captured in the *Psychiatric/Mental Health Rate* strata), based on evidence that the clinical needs for these patients substantively differ from those patients seeking non-psychiatric treatment (Nazarian et al. 2017).

Collectively, the findings from these studies and guideline suggest that there is room for improvement in the time from a patient's arrival to the time of his or her departure, and that important differences in both the ED throughput time and overall treatment approach exist for those seeking mental health or psychiatric treatment, when compared to the overall population. This evidence base supports the continued utility of NQF #0496.

1a.4.2 What process was used to identify the evidence?

2014 Submission: Environmental scan

2018 Submission: The measure developer conducted a review of clinical practice guidelines, peer-reviewed literature, and related policy during the NQF #0496's annual literature review to identify additional evidence and/or new studies that support the measure's intent. The measure developer identified relevant peer-reviewed publications by searching the PubMed MEDLINE database for evidence made available from January 1, 2013 to September 30, 2017, limiting included results to those published in the English language and that had abstracts available in PubMed. The search initially identified 127 articles; a further review by the developer's team refined this evidence base, resulting in the inclusion of seven articles in the body of evidence below. Citations and abstracts from this effort can be found in **Section 1a.4.3**.

1a.4.3. Provide the citation(s) for the evidence.

2014 Submission:

- Pines JM, Hollander JE, Localio AR, Metlay JP. The association between emergency department crowding and hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. Acad Emerg Med. 2006 Aug;13(8):873-8.
- Fee C, Weber EJ, Maak CA, Bacchetti P. Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. Ann Emerg Med. 2007 Nov; 50(5):501-9, 509.e1.
- Diercks DB, Roe MT, Chen AY, Peacock WF, Kirk JD, Pollack CV Jr, Gibler WB, Smith SC Jr, Ohman M, Peterson ED. Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. Ann Emerg Med. 2007 Nov;50(5):489-96.
- Chaflin DB, Trzeciak S, Likourezos A, et al. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. Crit Care Med. 2007;35:1477-1483.
- Richardson DB. Increase in patient mortality at 10 days associated with emergency department overcrowding. Med J Aust. 2006;184:213-216.
- Sprivulis PC, DaSilva JA, Jacobs IG, et al. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. Med J Aust. 2006;184:208-212.
- Carr BG, Kaye AJ, Wiebe DJ, et al. Emergency department length of stay: a major risk factor for pneumonia in intubated blunt trauma patients. J Trauma. 2007;63:9-12.
- Pines JM, Localio AR, Hollander JE. The impact of emergency department crowding measures on time to antibiotics for patients with community-acquired pneumonia. Ann Emerg Med. 2007;50: 510-516.
- Derlet RW, Richards JR. Emergency department overcrowding in Florida, New York, and Texas. South Med J. 2002;95:846-9.
- Derlet RW, Richards JR. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Ann Emerg Med. 2000;35:63-8.

- Fatovich DM, Hirsch RL. Entry overload, emergency department overcrowding, and ambulance bypass. Emerg Med J. 2003;20:406-9.
- Hwang U, Richardson LD, Sonuyi TO, Morrison RS. The effect of emergency department crowding on the management of pain in older adults with hip fracture. J Am Geriatr Soc. 2006;54:270-5.
- Krochmal P, Riley TA. Increased health care costs associated with ED overcrowding. Am J Emerg Med. 1994;12:265-6.
- Kyriacou DN, Ricketts V, Dyne PL, McCollough MD, Talan DA. A 5-year time study analysis of emergency department patient care efficiency. Ann Emerg Med. 1999;34:326-35.
- Nawar ED, Niska RW, Xu J. National Hospital Ambulatory Medical Care Survey: 2005 emergency department summary. Adv Data. 2007; (386):1-32.
- Richardson DB. Increase in patient mortality at 10 days associated with emergency department overcrowding. Med J Aust. 2006;184:213-6.
- Sprivulis PC, et al. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. Med J Aust. 2006;184:208-12.
- Trzeciak S, Rivers EP. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. Emerg Med J. 2003;20:402-5.
- Wilper AP, Woolhandler S, Lasser KE, McCormick D, Cutrona SL, Bor DH, Himmelstein DU. Waits to see an emergency department physician: U.S. trends and predictors, 1997-2004. Health Aff (Millwood). 2008;27:w84-95.

2018 Submission:

Bucci, S., A. G. de Belvis, S. Marventano, A. C. De Leva, M. Tanzariello, M. L. Specchia, W. Ricciardi and F. Franceschi. Emergency department crowding and hospital bed shortage: Is Lean a smart answer? A systematic review. Eur Rev Med Pharmacol Sci, 2016, 20(20), 4209-4219.

OBJECTIVE: Emergency Departments (EDs) worldwide face the challenges of crowding, waiting times, and cost containment. This review aims to provide a synthesis of the current literature focused on how Lean Thinking Principles and tools can be applied in an ED to address overcrowding and hospital admissions. MATERIALS AND METHODS: Primary studies showing Lean interventions and implementation in ED visits, not requiring additional resources measuring specific outcomes (i.e. length of stay, patient volume, patient satisfaction, waiting times for the first visit, waiting times for diagnostic results, left without being seen) were selected. PubMed, Scopus, CINAHL, EconLit, NHS Economic Evaluation Database, Business Sources Complete, and Health Technology Assessment were used to conduct searches. Full-text articles of all potentially relevant publications were reviewed for eligibility. Discrepancies were resolved through discussion by all reviewers. Quality assessment and critical appraisal of selected studies were also evaluated by applying the Quality Improvement Minimum Quality Criteria Set. RESULTS: Nine before-and-after studies met these eligibility criteria. Management of patient flow was the main intervention. Almost all studies showed EDs performance improvement: increased patient volume, decreased length of stay and number of patients left without being seen, reduced costs, and increased patient satisfaction. Only one case reported worse results after Lean intervention implementation. CONCLUSIONS: Though Lean Principals have been used in healthcare for many years conclusion of their effects could still not be drawn. Surely, human-centered approach, top management support, work standardization, resources allocation and adaptation to the local context seem to be crucial for success. Furthermore, higher quality studies are needed: specific research design, appropriate statistical tests and outcome measures are needed. Before large-scale implementation, further studies are needed to evaluate the true ability of Lean interventions to improve healthcare delivery.

Chang, A. M., A. Lin, R. Fu, K. J. McConnell and B. Sun. Associations of Emergency Department Length of Stay With Publicly Reported Quality-of-care Measures. Acad Emerg Med, 2017, 24(2), 246-250.

Chang et al. studied assessed the association between changes in publicly reported ED length of stay (LOS) and changes in quality-of-care measures in a national cohort of hospitals. The cohort consisted of 2,619 hospitals. Each additional hour of ED LOS was associated with a 0.7% decrease in proportion of patients giving a top satisfaction rating, a 0.7% decrease in proportion of patients who would “definitely recommend” the hospital, and a 6-minute increase in time to pain management for long bone fracture (p < 0.01 for all). A 1-hour increase in ED LOS is associated with a 44% increase in the odds of having an increase in left without being seen (95% confidence interval = 25% to 68%). ED LOS was not associated with hospital readmissions (p = 0.14) or time to percutaneous coronary intervention (p = 0.14). In this longitudinal study of hospitals across the United States, improvements in ED timeliness measures are associated with improvements in the patient experience.

Gardner, R. M., N. A. Friedman, M. Carlson, T. S. Bradham and T. W. Barrett. (2017). Impact of revised triage to improve throughput in an ED with limited traditional fast track population. Am J Emerg Med., 36(1), 124-127.

BACKGROUND: Emergency department (ED) crowding is associated with patient safety concerns, increased patients left without being seen (LWBS), low patient satisfaction, and lost ED revenue. The objective was to measure the impact of a revised triage process on ED throughput. METHODS: This study took place at an urban, university-affiliated, adult ED with an annual census of 70,000 and admission rate of 34%. The revised triage approach included: identifying eligible patients at triage based on complaint, comorbidities, and illness acuity; and reallocating a nurse practitioner (NP) into our triage area. We trialed the intervention from 1100-2300 on weekdays from January 13-26, 2016. Adult patients who were not likely to require intensive evaluations were eligible. Primary outcomes were throughput measures including: time to provider, ED length of stay (LOS), and LWBS. Pre- and post-intervention metrics were compared using the Mann-Whitney U test, given the non-normal distribution of the metrics. RESULTS: The NP evaluated 120 patients of which 101 (84%) were discharged, 3 (2.5%) admitted, and 16 (13%) required more intense evaluation. Time to provider decreased from a median (IQR) of 42 (16, 114) to 27 (12.4, 81.5) minutes (p<0.01) and ED LOS from 290 (194.8, 405.6) to 257 (171.2, 363.4) minutes (p<0.01) for all patients not admitted and not requiring a consult. LWBS decreased from a pre-trial 4.6% to 2.2% (p<0.01). CONCLUSION: The revised triage intervention was associated with improvements in several ED throughput metrics and a reduction in LWBS.

Mullins PM, Pines JM. National ED crowding and hospital quality: Results from the 2013 Hospital Compare data. Am J Emerg Med 2014; 32(6): 634-639.

We explored Hospital Compare data on emergency department (ED) crowding metrics to assess characteristics of reporting vs non-reporting hospitals, whether hospitals ranked as the US News Best Hospitals (2012-2013) vs unranked hospitals differed in ED performance and relationships between ED crowding and other reported hospital quality measures. An ecological study was conducted using data from Hospital Compare data sets released March 2013 and from a popular press publication, US News Best Hospitals 2012 to 2013. We compared hospitals on 5 ED crowding measures: left-without-being-seen rates, waiting times, boarding times, and length of stay for admitted and discharged patients. Of 4810 hospitals included in the Hospital Compare sample, 2990 (62.2%) reported all ED 5 crowding measures. Median ED length of stay for admitted patients was 262 minutes (interquartile range [IQR], 215-326), median boarding was 88 minutes (IQR, 60-128), median ED length of stay for discharged patients was 139 minutes (IQR, 114-168), and median waiting time was 30 minutes (IQR, 20-44). Hospitals ranked as US News Best Hospitals 2012 to 2013 (n=650) reported poorer performance on ED crowding measures than unranked hospitals (n=4160) across all measures. Emergency department boarding times were associated with readmission rates for acute myocardial infarction (r=0.14, P<.001) and pneumonia (r=0.17, P<.001) as well as central line-associated bloodstream infections (r=0.37, P<.001). There is great variation in measures of ED crowding across the United States. Emergency department crowding was related to several measures of in-patient quality, which suggests that ED crowding should be a hospital-wide priority for quality improvement efforts.

Nazarian DJ, Broder JS, Thiessen ME, Wilson MP, Zun LS, Brown MD, American College of Emergency Physicians. Clinical policy: critical issues in the diagnosis and management of the adult psychiatric patients in the emergency department. Ann Emerg Med. 2017 Apr; 69(4):480-98.

This clinical policy from the American College of Emergency Physicians addresses key issues for the diagnosis and management of adult psychiatric patients in the emergency department. A writing subcommittee conducted a systematic review of the literature to derive evidence-based recommendations to answer the following clinical questions: (1) In the alert adult patient presenting to the emergency department with acute psychiatric symptoms, should routine laboratory tests be used to identify contributory medical conditions (non-psychiatric disorders)? (2) In the adult patient with new-onset psychosis without focal neurologic deficit, should brain imaging be obtained acutely? (3) In the adult patient presenting to the emergency department with suicidal ideation, can risk-assessment tools in the emergency department identify those who are safe for discharge? (4) In the adult patient presenting to the emergency department with acute agitation, can ketamine be used safely and effectively? Evidence was graded and recommendations were made based on the strength of the available data.

Sun, B. C., A. Laurie, L. Prewitt, R. Fu, A. M. Chang, J. Augustine, C. t. Reese and K. J. McConnell. "Risk-Adjusted Variation of Publicly Reported Emergency Department Timeliness Measures." Annals of Emergency Medicine, 2016,67(4):509-516 e7.

The Centers for Medicare & Medicaid Services (CMS) recently published emergency department (ED) timeliness measures. These data show substantial variation in hospital performance and suggest the need for process improvement initiatives. However, the CMS measures are not risk adjusted and may provide misleading information about hospital performance and variation. We hypothesize that substantial hospital-level variation will persist after risk adjustment. This cross-sectional study included hospitals that participated in the Emergency Department Benchmarking Alliance and CMS ED measure reporting in 2012. Outcomes included the CMS measures corresponding to median annual boarding time, length of stay of admitted patients, length of stay of discharged patients, and waiting time of discharged patients. Covariates included hospital structural characteristics and case-mix information from the American Hospital Association Survey, CMS cost reports, and the Emergency Department Benchmarking Alliance. We used a γ regression with a log link to model the skewed outcomes. We used indirect standardization to create risk-adjusted measures. We defined "substantial" variation as coefficient of variation greater than 0.15. The study cohort included 723 hospitals. Risk-adjusted performance on the CMS measures varied substantially across hospitals, with coefficient of variation greater than 0.15 for all measures. Ratios between the 10th and 90th percentiles of performance ranged from 1.5-fold for length of stay of discharged patients to 3-fold for waiting time of discharged patients. Policy-relevant variations in publicly reported CMS ED timeliness measures persist after risk adjustment for nonmodifiable hospital and case-mix characteristics. Future "positive deviance" studies should identify modifiable process measures associated with high performance.

Zocchi, M. S., M. S. McClelland, and J. M. Pines. Increasing Throughput: Results From A 42-Hospital Collaborative To Improve Emergency Department Flow. The Joint Commission Journal on Quality and Patient Safety, 2015, 41(12):532–542.

BACKGROUND: An 18-month collaborative in 42 hospitals across 16 communities in the United States to improve emergency department (ED) flow was conducted from October 2010 through March 2012. METHODS: Hospitals were invited to participate through the Aligning Forces for Quality (AF4Q) program. Each participating hospital identified one or more interventions to improve ED flow and submitted data on four measures of ED flow: discharged length of stay (LOS), admitted LOS, boarding time, and left without being seen (LWBS) rates. Participating hospitals also provided quarterly progress reports on challenges encountered and lessons learned. Univariate linear regression was used to assess the effectiveness of interventions at the hospital level, where an improvement was defined as a negative slope in one or more of the throughput indicators. Challenges and lessons learned were tabulated and described. RESULTS: A total of 172 interventions were implemented across the 42 hospitals. Two thirds (n = 28) demonstrated improvement on at least one measure of ED flow. Among hospitals demonstrating improvement, the average reduction in discharged LOS was 26 minutes (95% confidence interval [CI] 11 to 41); admitted LOS, 36.5 minutes (95% CI 20 to 53), boarding time, 20.9 minutes (95% CI 12 to 30), and LWBS seen rates decreased by 1.4 absolute percentage points (95% CI 0.2 to 2.7). Teams were frequently challenged by issues related to leadership, staff buy-in, and resource constraints. CONCLUSION: The majority of hospitals in this collaborative improved on one or more ED flow measures. Many challenges were shared across hospitals, demonstrating that successful approaches to ED flow improvement require certain fundamental elements, including engaged leadership and staff, and sufficient resources.

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.
- 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] 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] 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]

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]

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

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

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

[1] 2014 testing information is included in Appendix B.

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*.

a) Datasets used to <u>define the sample:</u>
   - 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.

b) Datasets used to<u> define the effective sample</u> 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.

c) Datasets used to <u>identify denominator exclusions:</u>
   - 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 code related 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.

d) Datasets used to <u>identify numerator exceptions:</u>
   - 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.

e) 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:
a) *Denominator:* Clinical Data Warehouse (CDW), maintained by the Centers for Medicare & Medicaid Services (CMS)
b) *Numerator:* CDW
c) *Exclusions:* CDW
d) *Exceptions:* CDW

Dates:
a) *Denominator:* October 01, 2015–August 30, 2016
b) *Numerator:* October 01, 2015–August 30, 2016
c) *Exclusions:* October 01, 2015–August 30, 2016
d) *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]

Data source:
a) *Denominator:* CDW
b) *Numerator:* CDW
c) *Exclusions:* CDW
d) *Exceptions:* CDW

Dates:
a) *Denominator:* October 01, 2015–August 30, 2016
b) *Numerator:* October 01, 2015–August 30, 2016
c) *Exclusions:* October 01, 2015–August 30, 2016
d) *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

[2] CDAC is considered to be an authoritative source to which data from facility abstraction are compared.

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:
a) *Denominator:* CDW
b) *Numerator:* CDW
c) *Exclusions:* CDW
d) *Exceptions:* CDW

Dates:
a) *Denominator:* October 01, 2015–August 30, 2016
b) *Numerator:* October 01, 2015–August 30, 2016
c) *Exclusions:* October 01, 2015–August 30, 2016
d) *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 *overall*, *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:
a) *Denominator:* January 1, 2016–December 31, 2016
b) *Numerator:* January 1, 2016–December 31, 2016
c) *Exclusions:* January 1, 2016–December 31, 2016
d) *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:
a) *Denominator:* CDW

b) *Numerator:* CDW
c) *Exclusions:* CDW
d) *Exceptions:* CDW

Dates:
a) *Denominator:* October 01, 2015–August 30, 2016
b) *Numerator:* October 01, 2015–August 30, 2016
c) *Exclusions:* October 01, 2015–August 30, 2016
d) *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*. The reliability 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:

$$\text{ICC} = \frac{variance_{facility}}{variance_{facility} + variance_{error}}$$

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] 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*

[3] The 99th percentile is based on the measure score of cases included in the *Overall* rate.

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]	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 <u>performance measure score</u> 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

[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.

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 know*, 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

2. *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

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.*

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

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.*

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

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?*

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

6. *To be included in the NQF #0496 (OP-18)*[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

[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).

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*

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

Data Element	**Denominator Exclusion or Numerator Exception?**		**Overall Occurrence**		**Distribution across Facilities (%)**		
	Denominator Exclusion	*Numerator Exception*	*N*	*%*	*25th*	*50th*	*75th*
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]	X		4,835	0.2	0.0	0.0	0.2
Discharge Code equal to 4a or 4d [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.

[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.

[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.

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.

a) *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%.

b) *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.

c) *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.

d) *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.

e) *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.*

f) *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.

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

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 Statistics[a]/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	0.17[a]

a - Kappa Statistics

* - ICC