

Reliability, Encounter Level Testing and Measure Performance Score Results for the Emergency Care Capacity and Quality (ECCQ) Electronic Clinical Quality Measure (eCQM) for the Hospital Outpatient Quality Reporting Program

Please note that due to limitations with testing, reliability, encounter level testing and measure performance score results were not available by the May 2024 Measures Under Consideration (MUC) submission deadline. As noted within the MUC Entry/Review Information Tool (MERIT) submission, testing results would be available by August 2024. Therefore, the results are included within this attachment, using the format of the 2024 MERIT Data Template for the relevant sections below. We request that reviewers reference this attachment for final testing results.

Data Source

Three datasets were used to test the ECCQ eCQM, derived from electronic health record (EHR) data from multiple testing partners. There was an overall mix of geographic regions, hospital size, teaching status, trauma level, and EHR vendor. Dataset A was the primary dataset used for testing, with Dataset B used to replicate the measure scores and calculate reliability, and Dataset C used for data element validity.

Dataset A consisted of a diverse array of 20 emergency departments (EDs). Dataset represented 11 health systems, Epic and Cerner EHR systems, four rural EDs, and a mix of geographic locations, bed size, teaching status, and trauma level. Four sites were rural: ED4, ED7, ED14, ED19. All but ED19 are a part of a larger health system. For Dataset A we used calendar years 2022 and 2023, both combined and as separate performance periods, for different types of analyses. This allowed us to look at the measure score for each site over two years, creating 40 data points, to test volume standardization (comparing hospital scores with similar number of encounters), and it allowed us to see measure score changes year over year.

Results are labeled by the dataset name and year as follows:

- **Dataset A 2-years (2022-2023)** was 20 EDs over two years, so using 40 ED data points; it represented 2,196,714 encounters;
- **Dataset A 2022** had 20 EDs, representing 1,077,773 encounters; and
- **Dataset A 2023** had the same 20 EDs, representing 1,118,941 encounters.

Dataset A included all required data elements to calculate measure scores, and patient characteristics such as date of birth, gender/sex, race, payer.

Dataset B consisted of 12 hospital-based EDs, representing 832,056 encounters, within one large health system, using Epic EHR system. EDs were in the South, ranging in bed size, teaching status, and trauma level. Two EDs were rural: ED1, ED3. Dataset B included one calendar year of data, 2023.

Dataset B included all required data elements to calculate measure scores, and patient characteristics such as date of birth, gender/sex, and race, but not payer.

Dataset C consisted of six hospital-based EDs within one health system, using Epic EHR systems. EDs were in the Northeast, ranging in bed size, teaching status, with one trauma center. Dataset C includes calendar year 2023.

Reliability: Measure Score Level (Accountable Entity Level) Testing

Reliability testing was completed for this measure using a Signal-to-Noise calculation at the facility-level, using a beta binomial model.

32 accountable entities (hospital-based ED facilities) (Dataset A + B) were included in the testing.

Table 1. Measure Score Level Reliability Analysis: Signal-to-Noise, Dataset A 2023 and Dataset B (N=32 EDs, 2,740,383 encounters)

Signal-to-Noise Statistic	Mean (SD)	Range (min-max)
Dataset A 2023 and Dataset B combined (N=32), 2023	.9999 (0)	(.9997 – 1.000)

Interpretation of results:

Overall results show high reliability for all measured entities, indicating the measure is reliable and the wide range in measure scores are due to differences in quality.

Empiric Validity: Measure Score Level (Accountability Entity Level) Testing

To enhance the measure’s validity, we conducted an analysis of construct validity, the extent to which the measure accurately assesses what it is intended to. This was not originally planned for testing at the time of the MUC submission in May, however the opportunity for additional validity testing arose and the results are included in support of the measure’s empiric validity.

32 accountable entities (hospital-based ED facilities) from Dataset A 2023 and Dataset B overall measure scores were included in the testing, however not every analysis was performed combined with Datasets A and B. We tested the construct validity at the facility-level using with a Pearson’s correlation coefficient to examine the association between measure score performance and broadly available and validated measures of hospital quality including the Overall Hospital Quality Star Rating, the Hospital Quality Summary Score, and domain-level quality scores in mortality, readmission, patient experience, and timely care. We hypothesized a negative correlation with each of the Star Ratings components because hospital capacity, including ECCQ numerator components have been shown to be associated with hospital quality across a range of outcomes including mortality, patient experience, and cost.

The results in [Table 2](#) indicate the statistical results affirmed the hypothesized relationship; hospitals that performed well on Star Ratings also performed well on ECCQ eCQM, lending validity to the novel ECCQ eCQM.

Table 2. Correlation Between the ECCQ eCQM and Validated Existing Measures of Hospital Quality

Dataset	Expected Relationship	Overall Hospital Quality Star Rating	Hospital Quality Summary Score	Domain-Level Quality Score: Readmission	Domain-Level Quality Score: Mortality	Domain-Level Quality Score: Timely Care
Dataset A, 2022	Negative	-0.56	-0.73	-0.61	-0.13	-0.54
Dataset B, 2022	Negative	-0.55	-0.53	-0.51	-0.26	-0.35

Face Validity: Measure Score Level (Accountable Entity Level) Testing

We systematically assessed the face validity of the ECCQ eCQM for the Hospital Outpatient Quality Reporting (HOQR) Program measure score as an indicator of quality by soliciting the experts and patients/caregivers’ agreement with the following statement: “The Emergency Care Capacity and Quality eCQM for the HOQR Program could differentiate good from poor quality of care among facilities.” At the time of the face validity vote, the measure specifications included a numerator exclusion removing transfers to another facility from calculation in component #4 ED length of stay (LOS). CORE does not believe this greatly impacts the face validity; experts widely agreed upon the importance of transfers relative to the measure’s intent and importance.

A total of 16 Technical Expert Panel (TEP) members responded. The scale was as follows: *strongly agree, agree, disagree, strongly disagree*. Results of the TEP rating of agreement with the validity statement were as follows:

Specifically, 75.0% of TEP members agreed that the ECCQ eCQM measure could differentiate good from poor quality of care. There were 8 votes for strongly agree (50.0%), 4 votes for agree (25.0%), 4 votes for disagree (25.0%), and 0 votes for strongly disagree (0.0 %). Members who voted in agreement noted that these metrics are correlated with patient outcomes, so it is a useful quality measure with good face validity and construct validity. The measure considers various components that are proxies for access to emergency care, noting that a key tenet of emergency care is that it is timely, and this measure that can capture the data necessary to drive hospitals to improve throughput.

The 25.0% of TEP members who voted disagree noted that they disagreed that the measure could differentiate good from poor quality of care based on the boarding and ED LOS threshold, as the factors driving those are not exclusively within the facilities’ control. They disagreed in the definition of how a private treatment space is being defined and recorded. They noted the measure is of time, organizational capacity, and efficiency but not quality of care. Lastly, another member disagreed because the measure does not adjust for trauma levels designated to hospitals.

Patient/Encounter Level (Data Element Level) Testing

Encounter-level testing of the individual data elements in the final performance measure (i.e., measure of agreement between eCQM and manual reviewers, by percent agreement) was completed. We

assessed data element validity by the raw match rate of each required EHR data element to the chart abstracted data element. We validated the numerator events, denominator-only encounters, and those in the numerator exclusion (observation stays). We considered each data element “matched” if the electronically extracted value (from EHR) exactly matched the manual abstraction value (from the patient medical record).

Data element validity testing was conducted using a sample of 254 patient charts of ED encounters; the percent agreement was used to assess agreement between eCQM and manual chart review to demonstrate validity of the critical, required data elements. This sample included 20 observation stays, 20 transfers, 20 admitted patients, 10 left without being seen (LWBS) cases, 50 denominator-only cases, and 130 numerator cases.

Validation of ED encounters by disposition and data elements demonstrated high validity and high levels of agreement between electronic record review and manual chart review.

- 95% of admissions records were confirmed through manual chart review.
- 100% of transfer records, final ED dispositions, ED arrival time, and time placed in treatment room were confirmed through manual chart review.
- 37% of reviewed records (94 out of 254) had a documented admission time, indicating 94 patients were admitted to the hospital, and of those admitted, 100% of the records had an exact match of the inpatient admission timestamp.
- 96% of reviewed records had an exact match of ED departure timestamp (245 out of 254 records), for the 9 non-matching records:
 - 7 records had a discrepancy in time of less than 90 minutes.
 - 2 records were outliers, with a wide discrepancy in discharge time, and discrepancy caused by a readmission.

Further analyses explored the percent agreement of each timestamp used to calculate the data elements for each numerator component of the measure. The results show percentage agreement in [Table 3](#).

Table 3. Percent agreement of numerator components between eCQM and manual chart review

Numerator Component	Percent Agreement
Time to Placement in Waiting Room	100%
Left without being seen	100%
Boarding	100%
ED Length of Stay	100%
Any numerator	99.6%

Interpretation of Results

Data element validity requires high rates of data capture and low rates of missing data, and this analysis of validity supports that this ECCQ eCQM specification relies on available electronic time stamps that are routinely present in clinical notes and therefore the measure is best suited for electronic capture.

All available evidence indicates that data element reliability is high within one health system which supports the validity of this ECCQ eCQM; further testing may be required to demonstrate reliability and validity across more health systems.

Measure Performance

ECCQ eCQM is a proportion measure where better quality = lower score.

A total of 32 accountable entities (hospital-based ED facilities) were included in the analysis of the distribution of performance scores. The analyses in [Table 4](#) provide the measure scores for each site in Dataset A split by year (2022 and 2023) and stratified by age and mental health (MH) status. Dataset A had 20 hospitals with 2,196,714 ED encounters. Dataset B had 12 hospitals with 832,056 ED encounters.

Measure performance scores are proportion of numerator encounters, calculated as the number of encounters where any one of the four numerator criteria are met.

Mean, standard deviation (SD), median (interquartile range (IQR)), minimum and maximum are included in Table 3 below. 10th and 90th percentiles are not included due to few accountable entities this would not provide more detailed information.

Social risk factor impact on measure scores was not tested.

Table 4. Distribution of unadjusted measure scores in Dataset A and Dataset B

Measure Score	Mean (SD)	Median (IQR)	Range (min-max)
Dataset A 2022, 2023			
EDs Overall (N=40)	26.60 (16.07)	30.36 (10.36-39.96)	(2.91-55.91)
EDs Entire Cohort, 2022 (N=20)	28.28 (16.63)	34.28 (10.83-39.83)	(3.52-55.91)
EDs Entire Cohort, 2023 (N=20)	24.92 (15.75)	26.30 (10.36-40.19)	(2.91-52.13)
Adult Non-Mental Health Strata (N=20)	28.02 (17.01)	32.47 (10.84-40.59)	(3.68-59.53)
Adult Mental Health Strata (N=20)	32.67 (19.85)	29.60 (14.78-45.91)	(8.52-70.80)
Pediatric Non-Mental Health Strata (N=20)	18.22 (12.50)	15.28 (8.94-27.36)	(1.61-40.73)
Pediatric Mental Health Strata (N=20)	22.90 (12.08)	20.54 (13.74-32.06)	(2.75-50.00)
Dataset B 2023			
EDs Entire Cohort (N=12)	23.87 (5.36)	24.07 (20.28-27.97)	(15.91-32.21)
Adult Non-Mental Health Strata (N=12)	23.59 (4.82)	23.54 (20.23-27.30)	(15.90-30.90)
Adult Mental Health Strata (N=12)	49.93 (10.55)	52.27 (41.35-57.57)	(34.57-66.48)
Pediatric Non-Mental Health Strata (N=12)	16.67 (10.15)	14.94 (10.04-24.37)	(2.98-34.07)
Pediatric Mental Health Strata (N=12)	52.62 (10.89)	52.19 (46.59-58.54)	(33.82-71.62)

Additionally, below is the distribution (mean, standard deviation, median, minimum and maximum) of numerator components by strata, indicating which component had the potential to trigger the numerator (outcome), in [Table 5](#); an ED encounter with any one of the numerator components triggered will be included in the numerator (outcome). While each encounter is only counted once in the measure numerator, the numerator components displayed in this table (Numerators 1-4) are counted individually

and are not mutually exclusive (one encounter may include triggering of more than one of the four components) in this table. The strata are mutually exclusive (i.e., each encounter is counted in only one strata). This analysis used Dataset A 2023.

Not every ED had patients that triggered every numerator criterion. For example, two EDs did not have any of their patients under 18 with mental health diagnosis wait longer than one hour to be treated. In contrast, every ED had some adult patients who waited longer than one hour, and some pediatric patients without mental health diagnosis who waited longer than one hour.

Table 5. Distribution of Numerator Components per strata, Dataset A 2023

Numerator Component	Strata	# EDs	Mean (SD)	Median (IQR)	Range (min-max)
Numerator 1 (Wait time 1+ hour)	18+ Non MH	20	14.01 (9.62)	14.44 (5.58-20.66)	(0.39-31.86)
	<18 Non MH	20	13.57 (9.24)	11.94 (6.48-22.15)	(0.52-30.00)
	18+ MH	20	7.87 (7.31)	4.92 (2.82-12.32)	(0.26-29.54)
	<18 MH	18	8.83 (10.36)	4.99 (1.78-11.92)	(0.27-34.14)
Numerator 2 (LWBS)	18+ Non MH	20	2.60 (2.50)	1.62 (0.78-4.28)	(0.21-10.15)
	<18 Non MH	20	3.28 (3.44)	2.27 (0.74-3.78)	(0.21-13.31)
	18+ MH	13	0.26 (0.28)	0.13 (0.09-0.30)	(0.05-0.97)
	<18 MH	8	0.22 (0.17)	0.16 (0.13-0.24)	(0.06-0.60)
Numerator 3 (Boarded 4+ hours)	18+ Non MH	20	5.96 (4.38)	5.43 (1.88-10.11)	(0.18-12.75)
	<18 Non MH	18	0.48 (0.64)	0.24 (0.07-0.37)	(0.01-2.05)
	18+ MH	20	3.23 (4.21)	1.99 (1.05-3.60)	(0.19-18.80)
	<18 MH	13	1.43 (3.23)	0.46 (0.27-0.71)	(0.15-12.07)
Numerator 4 (LOS 8+ hours)	18+ Non MH	20	15.08 (10.30)	14.08 (5.36-25.57)	(2.33-33.51)
	<18 Non MH	20	2.54 (1.70)	2.25 (1.08-4.34)	(0.29-6.40)
	18+ MH	20	41.86 (14.87)	38.95 (32.27-56.39)	(14.22-70.51)
	<18 MH	20	32.62 (12.19)	27.88 (23.71-43.74)	(14.68-60.56)

Interpretation of results:

There is a wide range in overall measure performance scores across both Datasets (2.9% - 70.80%) and across all strata which would indicate differential between hospitals and therefore room for quality improvement. Furthermore, there is a wide range of facility-level results across each numerator

component, which in part speaks to the variation in clinical processes by patient group and supports the measure performance and measure stratification.

Importance: Description of input collected from patients/caregivers:

Unfortunately, we had unexpectedly low in-person attendance at our patient workgroup meeting; the two patients/caregivers from the work group responded that they agreed that the measure is meaningful and produces information that is valuable in making care decisions.

The following question was asked to the two patient/caregivers consulted about the final measure: “Information from the Emergency Care Capacity and Quality eCQM is meaningful and produces information that is valuable in making care decisions.”

Patients/caregivers provided their response using a scale of *strongly agree, agree, disagree, strongly disagree*.

100% of patients/caregivers consulted either strongly agreed or agreed that the measure is meaningful and produces information that is valuable in making care decisions. One patient/caregivers responded, “strongly agree” and one patient/caregivers responded “agree”.