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

**Measure Number** (*if previously endorsed*)**:** 2363

**Measure Title**: Hospital Harm – Severe Hypoglycemia

**Date of Submission**: TBD

**Type of Measure:**

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

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

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

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

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

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

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

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

We acquired data from a patient safety organization to support alpha testing of the measure concept, data elements, and validity. We partnered with two health systems to complete beta testing of the MAT output in two different EHR systems. We assessed data element and measure score validity as well as measure score reliability in beta testing. The dataset used varies by testing type; see Section 1.7 for details.

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

The dates vary by testing type; see Section 1.7 for details.

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

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

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

The number of measured entities (hospitals) varies; see Section 1.7 for details.

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The number of admissions/patients varies; 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**.

The datasets, dates, number of measured hospitals, and number of admissions used in each phase of testing are in **Table 1**.

Table 1. Dataset Descriptions

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Applicable Section in the Testing Attachment** | **Description of Dataset** | **EHR Vendor** |
| **Beta Dataset 1** | Section 2a2 Reliability Testing  Section 2b1 Validity Testing  Section 2b4 Identification of Statistically Significant and Meaningful Differences in Performance  Section 2b6 Missing Data Analysis | Dates of Data: January 1, 2017 - December 31, 2017  Number of Hospitals: 4  Number of Admissions: 7,748  Number of Unique Patients: 5,394  For Validity Testing: sample of 175 admissions  Hospitals were within one health system, in urban locations. Hospitals ranged between 50 – 1,000 beds. Located in the Midwest. | Epic |
| **Beta Dataset 2** | Section 2a2 Reliability Testing  Section 2b1 Validity Testing  Section 2b4 Identification of Statistically Significant and Meaningful Differences in Performance  Section 2b6 Missing Data Analysis | Dates of Data: January 1, 2017 - December 31, 2017  Number of Hospitals: 2  Number of Admissions: 5,888  Number of Unique Patients: 4,580  For Validity Testing: sample of 175 admissions  Hospitals were within one health system, in urban locations. Hospitals ranged between 200 – 3,800 beds. Located in the South. | Cerner |
| **Alpha Dataset** | Section 2b1 Validity Testing (Measure Score) | Dates of Data: June 1, 2016 - May 31, 2017  Number of Hospitals: 5  Number of Admissions: 66,127  Hospitals were in two different health systems, both in urban locations, and not-for-profit. They were diverse in terms of bed size (between 100-199 beds and 300-399 bed), teaching status, geographic location (South, West). | Cerner & Epic |

**Patient descriptive characteristics in Alpha Dataset are as follows:**

* Patient Descriptive Characteristics:
  + Mean age at admission = 58.7 years with a standard deviation of 20.4 years
  + 58.2% female, 41.8% male
  + 64.5% White, 9.7% Black or African-American, 8.0% Asian, 1.0% Native Hawaiian or Other Pacific Islander, 0.2% American Indian or Alaska Native, 15.7% Other, and 0.9% declined or unknown

Patient descriptive characteristics included in the analysis by hospital for **Beta Datasets 1 and 2** are provided in **Table 2**.

Table 2. Demographic Characteristics of Eligible Patient Population (Beta Datasets 1 and 2)

| **Initial Patient Population Characteristics** | **Beta Dataset 1 (N, %)** | **Beta Dataset 2 (N, %)** | **Across Beta Sites (N, %)** |
| --- | --- | --- | --- |
| **Number of unique patients** | 5394 | 4580 | 9974 |
| **Average Age [Mean(SD)]** | 59 (15) | 65 (16) | 62 (15) |
| 18-35 | 415, 7.7% | 244, 5.3% | 659, 6.6% |
| 36-64 | 2929, 54.3% | 1827, 39.9% | 4756, 47.7% |
| 65+ | 2050, 38.0% | 2509, 54.8% | 4559, 45.7% |
| **Sex** |  |  |  |
| Male | 2930, 54.3% | 2304, 50.3% | 5234, 52.5% |
| Female | 2464, 45.7% | 2263, 49.4% | 4727, 47.4% |
| Unknown | 0, 0.0% | 13, 0.3% | 13, 0.1% |
| **Race** |  |  |  |
| Black or African-American | 1377, 25.5% | 638, 13.9% | 2015, 20.2% |
| White | 3729, 69.1% | 2458, 53.7% | 6187, 62.0% |
| Other | 288, 5.3% | 1269, 27.7% | 1557, 15.6% |
| Unknown | 0, 0.0% | 215, 4.7% | 215, 2.2% |
| **Ethnicity** |  |  |  |
| Hispanic or Latino | 82, 1.5% | 781, 17.1% | 863, 8.7% |
| Non-Hispanic | 5279, 97.9% | 3582, 78.2% | 8861, 88.8% |
| Unknown | 33, 0.6% | 217, 4.7% | 250, 2.5% |
| **(Primary) Payer** |  |  |  |
| Medicare | 2826, 52.4% | 2276, 49.7% | 5102, 51.2% |
| Medicaid | 572, 10.6% | 412, 9.0% | 984, 9.9% |
| Private Insurance | 1732, 32.1% | 1451, 31.7% | 3183, 31.9% |
| Self-pay or Uninsured | 4, 0.1% | 131, 2.9% | 135, 1.4% |
| Other (such as other government plans) | 178, 3.3% | 310, 6.8% | 488, 4.9% |
| Unknown | 82, 1.5% | 0, 0.0% | 82, 0.8% |

+ “Others” include all possible payers other than Medicare and Medicaid, such as other government plans (e.g. federal, state, local), private health insurance, etc.

**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.

As described in Section 1.7, Table 2, we collected information on the following social risk factors using data extracted from hospital EHR systems: race, ethnicity, and primary payer.

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

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “**see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

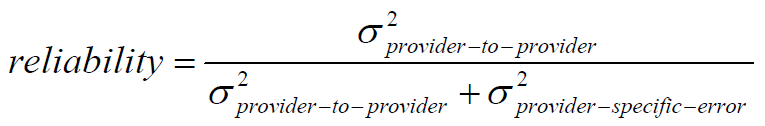
N/A. Since data element validity was empirically tested, separate reliability testing of data elements is not required per the NQF Measure Evaluation Criteria and Guidance (see section 2b2 for validity testing of data elements).

Measure Score Reliability

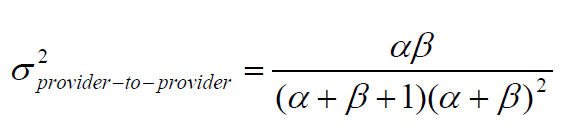
The reliability of a measure score is the degree to which repeated measurements of the same entity agree with each other. We estimated the measure score reliability using **Beta Datasets 1 and 2**.

We assessed signal- to-noise reliability that describes how well the measure can distinguish the performance of one hospital from another (Adams and Mehrota, 2010; Yu and Mehrota, 2013). The signal is the proportion of the variability in measured performance that can be explained by real differences in performance. Scores can range from 0 to 1. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance.

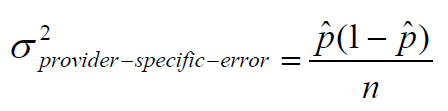
We use the Adam’s beta-binomial method (Adams, 2009) to calculate the signal-to-noise ratio reliability. Briefly, using variability between hospitals (signal: provider-to-provider variance) and variability within hospitals (noise: provider-specific-error variance), the reliability for each hospital can be defined as



We estimate the beta-binomial variance as the provider-to-provider variance as



where α, β are the estimated beta-binomial parameters using denominators and rates from all hospitals. The provider-specific-error variance is estimated as



where n is the numerator of a hospital and p ̂ is the harm rate of a hospital.

References:

Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Yu, H, Mehrota, A, Adams J. (2013). Reliability of utilization measures for primary care physician profiling. Healthcare, 1, 22-29.

Adams, J. The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corporation, 2009. https://www.rand.org/pubs/technical\_reports/TR653.html.

**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*)

Measure Score Reliability Results

There were 13,636 eligible encounters across 6 hospitals in **Beta Datasets 1 and 2**. The signal-to-noise ratio yielded a median reliability score of 0.889 (range: 0.815-0.924).

**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?*)

The signal-to-noise ratio of 0.89 indicates excellent agreement.

Our interpretation of these results is based on the standards established by Landis and Koch (1977):

< 0 – Less than chance agreement;

0 – 0.2 Slight agreement;

0.21 – 0.39 Fair agreement;

0.4 – 0.59 Moderate agreement;

0.6 – 0.79 Substantial agreement;

0.8 – 0.99 Almost Perfect agreement; and

1 Perfect agreement

Reference:

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Data element validity was assessed by evaluating the accuracy of electronically extracted EHR data elements compared with manually chart abstracted data elements from the same patients, which is considered the “gold standard” for the purpose of these analyses.

*Data Element Validity*:

For **Beta Datasets 1 and 2**, a stratified sample of 175 total admissions were selected at each hospital test site. Sample size calculations ensure a robust sample was used for validity testing. Specifically, we derived our sample size based on the following assumptions: Our primary endpoint for sample size estimation is PPV, which is applicable for both data element validity and measure score validity. We adjudicated all our numerator cases in alpha test and obtained high PPVs (>90% in most of the cases). Based on this, we approximate the sample size based on one-sample proportion formula as the following:

n=(moe/z\_(α/2) )^2\* p\* (1-p)

Where *a* is the type I error rate, *moe* is the margin of error, p is the proportion, here PPV, of interest. We simulate a series of *moe* and target PPV values for sample size and 95% confidence interval (CI) estimation. For example, with a *moe* of 6% and a target PPV of 0.9, a sample size of 100 will give rise to a 95% CI of 0.84 – 0.96. We concluded that a sample size of 100 from each hospital would ensure an accurate PPV estimation. Also, combining the samples from more than 1 hospitals would give us even more accurate estimation.

**Beta Dataset 1** had 175 encounters, 97 being admissions with harm events and 78 being admissions without a harm event (denominator-only); and **Beta Dataset 2** had 175 encounters, 100 being admissions with harm events and 75 being admissions without a harm event (denominator-only). Data were abstracted from the EHR by trained abstractors at each test site; abstractors at all sites had experience abstracting data for chart-based quality measure reporting. Abstractors were provided with an instruction manual and an Access database to document the information abstracted from the EHR. Access databases were only pre-populated with the unique patient identifier; abstractors were asked to input all other data from the chart independently of the EHR dataset. Abstraction training was also provided to each site.

**Table 3** shows the sensitivity agreement rate (# exact matches in both data sources / # sampled in the chart) between the data extracted from the EHR electronically and manual chart abstraction in **Beta Datasets 1 and 2**. Each data element matched if the specific electronically extracted value exactly matched the manually abstracted value (gold standard). For example, out of 84 specific instances where a patient was administered a antihyperglycemic medication (in the chart data), 68 of those specific cases were extracted correctly in the EHR data, resulting in an 81% match rate. For date/time data elements, we matched month, day, year, hour, and minutes. For glucose lab values, we matched on the glucose value result (whole integers), date, time +/- one minute. For administration of an antihyperglycemic medications, we matched on the name of the medication ordered, as the timestamps of medication administered in the EHR were autogenerated, and not found as easily in the chart.

Empirical Measure Score Validity

Measure score validity assesses whether the harm rate (or, the measure score outcome) calculated for each facility is in fact accurate. The measure score is calculated for each facility based on the number of encounters that experienced a harm compared to the total number of encounters. Therefore, we validated each individual harm identified in a sample of cases in the EHR by chart review by trained abstractors to confirm that the chart, or gold standard, reflects that a harm occurred. Because no further calculations are conducted to generate a facility level score (as is with risk-adjusted measures), We did not compare the harm rate to any other external measure of quality. For measures that count harm events without other statistical manipulation, the confirmation that the measure logic is accurately capturing true harm events is the gold standard for assessing validity of the measure score.

Therefore, to validate the EHR-extracted numerator against the gold standard of the patient medical chart, to assess whether the harms actually occurred and captured the intended outcome, we clinically adjudicated each admission that met the criteria for a harm among the sample of abstracted records, and calculated the positive predictive value (PPV) for all numerator cases and denominator cases, as shown in **Table 5,** in **Alpha Dataset, and Beta Datasets 1 and 2**. The PPV describes the probability that a patient with a positive result (numerator case) in the EHR data also was a positive result in the abstracted medical record data, as confirmed by a clinical adjudicator. Similarly, for denominator cases, the PPV describes the probability that a patient that was identified as a denominator case in the EHR was also a denominator case in the chart abstracted medical record data.

We also calculated the sensitivity, specificity, kappa, and negative predictive value (NPV) as shown in **Table 4** for Beta Dataset 1 and 2. Sensitivity describes the probability that a patient with a positive result in the abstracted medical record data was also a positive result in the EHR data. Specificity describes the probability that a patient with a negative result (not a numerator case) in the abstracted medical record data was also a negative result in the EHR data. Kappa describes the amount of remaining agreement between the harm incidences based on EHR and the harm incidences based on the abstracted medical record after the agreement by chance is taken into account. NPV describes the probability that a patient with a negative result (not in the numerator) in the EHR data also was a negative result in the abstracted medical record, confirmed by the clinical adjudicator.

For **Alpha Dataset**, data were abstracted from the EHR by trained abstractors who had experience abstracting data for chart-based quality measure reporting. Abstractors were provided with an instruction manual and an Excel, to document the information abstracted from the EHR. Abstraction training was also provided. Validity was established in the **Beta Datasets 1** **and 2** as described above.

Face Validity:

To systematically assess face validity, we surveyed our Technical Expert Panel (TEP), which is comprised of national experts and stakeholder organizations. We asked each member to rate the following statement using a six-point scale (1=Strongly Disagree, 2=Moderately Disagree, 3=Somewhat Disagree, 4=Somewhat Agree, 5= Moderately Agree, and 6=Strongly Agree): “the proportion of severe hypoglycemic events obtained from the Hospital Harm – Severe Hypoglycemia Measure as specified can be used to distinguish between better and worse quality care at hospitals.”

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

Data Element Validity

Table 3. Data Element Validity (Sensitivity) Results Required for Measure (Beta Datasets 1 and 2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Beta Dataset 1** | | | **Beta Dataset 2** | | |
| **# Cases Matched in EHR (n)** | **# Cases in Abstraction (n)** | **Sensitivity Percent Match (%)** | **# Cases Matched in EHR (n)** | **# Cases in Abstraction (n)** | **Sensitivity Percent Match (%)** |
| Admission date and time (mm/dd/yyyy, hh:mm) | 175 | 175 | 100.0% | 175 | 175 | 100.0% |
| Antihyperglycemic medication administered: order ID | 68 | 84 | 81.0% | 40 | 41 | 97.6% |
| Laboratory test, blood glucose test with date, time, result (mm/dd/yyy hh:mm XXX) | 2118 | 2215 | 95.6% | 2341 | 2454\* | 95.4% |
| Patient characteristic: birth date (mm/dd/yyyy) | 175 | 168 | 96.0%\*\* | 175 | 175 | 100.0%\*\* |

\*Data element validity for glucose result in **Beta Dataset 2** were matched using point-of-care tests only, due to human error by abstractor.

\*\*Patient date of birth was assessed using PPV percent match, as # cases matched in abstraction / # cases in EHR.

Empirical Measure Score Validity

**Table 4** displays the specificity, sensitivity, kappa, and NPV in each Beta Dataset. **Table 5** displays the positive predictive value (PPV) in each dataset. This PPV represents the percent of admissions that met the criteria for a harm (numerator) in the EHR confirmed by the chart abstraction, validated by a trained clinical adjudicator. **Alpha Dataset** validated the numerator cases and not denominator cases, due to data limitations. **Beta Datasets 1 and 2** were able to validate both numerator and denominator.

**Table 4. Measure Score Validity Statistics for Sample Between Electronic EHR Extraction and Manual Chart Abstraction (Sensitivity, Specificity, NPV, Kappa) (Bata Datasets 1 and 2)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Beta Dataset 1** | | | | **Beta Dataset 2** | | | |
| **Sensitivity** | **Specificity** | **Kapa (95% CI)** | **NPV** | **Sensitivity** | **Specificity** | **Kappa (95% CI)** | **NPV** |
| Severe Hypoglycemia | 100% | 95% | 0.95 (0.91, 1.0) | 100% | 100% | 94% | 0.94 (0.89, 1.0) | 100% |

Table 5. Measure Score Validity Statistics for Sample Between Electronic EHR Extraction and Manual Chart Abstraction (PPV) (Alpha Dataset, Beta Datasets 1 and 2)

|  |  |  |  |
| --- | --- | --- | --- |
| **Measure Component** | **Alpha Dataset PPV** | **Beta Dataset 1 PPV** | **Beta Dataset 2 PPV** |
|
| Initial patient population/ Denominator | N/A | 100.0% | 98.9% |
| Numerator | 99.2% | 95.9% | 95.0% |

Face Validity

*10 out of 11 TEP members responded to the survey question as follows: Moderately Disagreed (1), Somewhat Disagreed (1), Somewhat Agreed (0), Moderately Agreed (5), and Strongly Agreed (3).* The two TEP members who disagreed somewhat and moderately did not provide rationale.

**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?*)

Data Element Validity

All but one data element (at one site) had a match rate over 95%, indicating valid and accurate data elements were extracted from the EHR. The exception was for antihyperglycemic medication (81%) administered in **Beta Dataset 1**. We believe this specific match rate was due to different naming conventions between the way medications names are stored in the EHR and a basic drop-down list used by the chart abstracted data.

For the blood glucose date, time, and result data element, we assessed the validity of all glucoses recorded during the hospitalization, for a more robust sample to evaluate a clearer picture of data element accuracy. The match rate is below 100% due to the analysis matching on all three fields (date, time and result), which is a higher standard than matching of individual elements, and the timing of each variables does not match exactly within one minute. However, match rates were still high at 95.6% and 95.4% respectively. Overall, we believe the data elements required for the measure show validity.

Empirical Measure Score Validity

All three datasets had a PPV over 95%, meaning that almost all cases, the admission met the criteria for a harm in both the chart abstracted and EHR-extracted data. Although we do not always expect perfect agreement, as we expect some degree of human error in entering and matching values, we consider the PPV to show excellent measure score validity. The absence of a perfect PPV does not threaten validity as we do not expect any systematic error in this small amount of disagreement across hospitals that might bias the measure results. Similarly, specificity and sensitivity are high. Sensitivity is 100% in both Beta Dataset 1 and 2 and specificity is 95% and 94% in Beta Dataset 1 and 2 respectively. This means that the probability of the EHR data detecting a true hypoglycemic event in patients that had a true hypoglycemic event based on the abstracted data ('gold standard') is 100% (sensitivity). The probability of the EHR data detecting no hypoglycemia out of the no hypoglycemic event based on abstracted data is 94-95% (specificity). NPV was 100% in both Beta Dataset 1 and 2, indicating the EHR data indicated a harm did not occur, and 100% of the time the chart abstraction confirmed a harm did not occur. Kappa of 0.94 and 0.95 indicate excellent agreement. We will continue to reevaluate validity through reevaluation as hospitals participate in this measure and as required by NQF for maintenance of endorsement.

Our Kappa interpretation is based on the following standards set by Viera et al.:

0.4 – 0.6 indicate “moderate agreement”,

0.6 – 0.8 “substantial agreement”, and

0.8 – 1 “almost perfect agreement”

References:

1. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.

2. Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med 2005;37(5):360-3.

Face Validity:

80% of TEP members agreed (moderately or strongly) that the measure will provide an accurate reflection of quality, which reflects good face validity.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

**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*)

**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*)

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b4***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**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**.

Clinical characteristics, including a patient’s age, reason for hospitalization, clinical status when they arrive at the hospital, or comorbid conditions all may influence the risk of harm occurring during a hospitalization. Therefore, if hospitals care for patients with different degree of risk, it may be important to adjust for patient risk factors in order to compare hospital performance.

However, many harms such as severe hypoglycemia should be avoidable, regardless of patient risk. We consider the following criteria in determining whether risk adjustment is warranted for the severe hypoglycemia measure:

1. If many patients are at risk of the harm regardless of their age, clinical status, comorbidities, or reason for admission, as described further in paragraph below;
2. If the majority of incidents of the harm are linkable to care provision under the control of providers, for example harms caused by excessive or inappropriate medication dosing or inadequate monitoring; and
3. If there is evidence that the risk of a harm can be largely ameliorated by best care practices regardless of a patients’ inherent risk profile. For example, there may be evidence that even complex patients with multiple risk factors can avoid harm events when providers closely adhere to care guidelines

In the case of the severe hypoglycemia eCQM, there is evidence indicating that most hypoglycemic events of this severity (<40 mg/DL) are avoidable. Although certain patients may be particularly vulnerable to hypoglycemia in certain settings (e.g. due to organ failure and not related to administration of diabetic agents), the most common causes are lack of caloric intake, overuse of anti-diabetic agents, or both. As these causes are controllable in hospital environments, and risk can easily be reduced by following best practices, we do not think risk adjustment is warranted for this measure. We will continue to evaluate the appropriateness of risk adjustment in measure reevaluation as is required for NQF endorsement maintenance.

In addition to the clinical rationale provided for not risk adjusting this measure, we examined the performance (harm) rate of the measure across patient characteristics of age, sex, race, ethnicity, and payer. Age (by date of birth) was validated; no other patient demographic was validated using chart data. It is important to note these results are derived from a small dataset that is not generalizable to the entire population and the datasets include many characteristics that are ‘unknown’ in the EHR which limits the usability of the results; additionally, we do not believe it is clinically appropriate to adjust by these characteristics given the clinical rationale provided above.

**Table 6. Performance Rate by Encounter Characteristic (Beta Dataset 1 and 2)**

| **Characteristic** | **Beta Dataset 1** | | | **Beta Dataset 2** | | | | **Across Beta Sites** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Denominator** | **Numerator** | **Performance Rate % (95% CI)** | | **Denominator** | **Numerator** | **Performance Rate % (95% CI)** | **Denominator** | **Numerator** | **Performance Rate % (95% CI)** |
| **Number of unique Encounters** | 7,748 | 195 | 2.5 (2.2, 2.9) | | 5,888 | 174 | 3.0 (2.5, 3.4) | 13,636 | 369 | 2.71 (2.4, 3.0) |
| **Average Age** |  |  |  | |  |  |  |  |  |  |
| 18-64 | 4,882 | 119 | 2.4 (2.0, 2.9) | | 2,647 | 87 | 3.3 (2.6, 4.0) | 7,529 | 206 | 2.7 (2.4, 3.1) |
| 65+ | 2,866 | 76 | 2.7 (2.1, 3.3) | | 3,241 | 87 | 2.7 (2.2, 3.3) | 6,107 | 163 | 2.7 (2.3, 3.1) |
| **Sex** |  |  |  | |  |  |  |  |  |  |
| Male | 4,202 | 112 | 2.7 (2.2, 3.2) | | 2,928 | 85 | 2.9 (2.3, 3.6) | 7,130 | 197 | 2.8 (2.4, 3.2) |
| Female | 3,546 | 83 | 2.3 (1.9, 2.9) | | 2,941 | 87 | 3.0 (2.4, 3.6) | 6,487 | 170 | 2.6 (2.3, 3.0) |
| Unknown | 0 | 0 | N/A | | 19 | 2 | 10.6 (1.3, 33.1) | 19 | 2 | 10.5 (1.3, 33.1) |
| **Race** |  |  |  | |  |  |  |  |  |  |
| Black or African-American | 2,158 | 80 | 3.7 (3.0, 4.6) | | 809 | 34 | 4.2 (2.9, 5.8) | 2,967 | 114 | 3.8 (3.2, 4.6) |
| White | 5,193 | 104 | 2.0 (1.6, 2.4) | | 3,193 | 84 | 2.6 (2.1, 3.3) | 8,386 | 188 | 2.2 (2.0, 2.6) |
| Other | 397 | 11 | 2.8 (1.4, 4.9) | | 1,614 | 44 | 2.7 (2.0, 3.6) | 2,011 | 55 | 2.7 (2.1, 3.6) |
| Unknown | 0 | 0 | N/A | | 272 | 12 | 4.4 (2.3, 7.6) | 272 | 12 | 4.4 (2.3, 7.6) |
| **Ethnicity** |  |  |  | |  |  |  |  |  |  |
| Hispanic or Latino | 112 | 2 | 1.8 (0.2, 6.3) | | 968 | 28 | 2.9 (1.9, 4.2) | 1,080 | 30 | 2.8 (1.9, 3.9) |
| Non-Hispanic | 7,599 | 193 | 2.5 (2.2, 2.9) | | 4,602 | 137 | 3.0 (2.5, 3.5) | 12,201 | 330 | 2.7 (2.4, 3.0) |
| Unknown | 37 | 0 | 0.0 (0.0, 0.0) | | 318 | 9 | 2.8 (1.3, 5.3) | 355 | 9 | 2.5 (1.2, 4.8) |
| **(Primary) Payer** |  |  |  | |  |  |  |  |  |  |
| Medicare | 4,145 | 100 | 2.4 (2.0, 2.9) | | 3,016 | 92 | 3.1 (2.5, 3.7) | 7,161 | 192 | 2.7 (2.3, 3.1) |
| Medicaid | 792 | 23 | 2.9 (1.9, 4.3) | | 567 | 23 | 4.1 (2.6, 6.0) | 1,359 | 46 | 3.4 (2.5, 4.5) |
| Private Insurance | 2,468 | 68 | 2.8 (2.2, 3.5) | | 1,757 | 42 | 2.4 (1.7, 3.2) | 4,225 | 110 | 2.6 (2.1, 3.1) |
| Self-pay or Uninsured | 5 | 0 | 0.0 (0.0, 52.2) | | 166 | 3 | 1.8 (0.4, 5.2) | 171 | 3 | 1.8 (0.4, 5.0) |
| Other (such as other government plans) | 235 | 3 | 1.3 (0.26, 3.7) | | 382 | 14 | 3.7 (2.0, 6.1) | 617 | 17 | 2.8 (1.6, 4.4) |
| Unknown | 103 | 1 | 1.0 (0.02, 5.3) | | 0 | 0 | N/A | 103 | 1 | 1.0 (0.02, 5.3) |

**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?

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

**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.**

**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*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

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

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

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

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

**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*)

**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*)

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

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

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

We examined the data to determine if there were meaningful differences in performance (harm rates) between measured entities (i.e., hospitals). We examined confidence intervals around the estimates and variation in performance rates between hospitals within **Beta Datasets 1** **and 2** to determine the stability of each estimate and if there were differences in performance (harm rates) between hospitals, respectively.

**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*)

The performance rate across all hospitals in both **Beta Datasets 1 and 2** was 2.71% (95% CI: 2.44%, 2.99%). The performance rate ranged from 1.05% to 3.56% across all hospitals in both datasets.

The performance rate for all hospitals in **Beta Dataset 1** was 2.52% (95% CI: 2.18%, 2.89%).

The performance rate for all hospitals in **Beta Dataset 2** was 2.96% (95% CI: 2.54%, 3.42%).

**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?*)

Results from **Beta** **Datasets 1 and 2** showed performance scores that were within range of harm rates found in the literature (Nirantharakumar and Marshall, 2012; Wexler and Meigs, 2007). There was variation shown in the rate of harm across the six hospitals in this dataset, demonstrating a quality signal, suggesting room for improvement in rates of severe hypoglycemia among admitted patients.

References:

Nirantharakumar, K., Marshall, T., Kennedy, A., Narendran, P., Hemming, K., & Coleman, J. J. (2012). Hypoglycaemia is associated

with increased length of stay and mortality in people with diabetes who are hospitalized. Diabetic Medicine, 29(12), e445-e448.

Wexler, D. J., Meigs, J. B., Cagliero, E., Nathan, D. M., & Grant, R. W. (2007). Prevalence of hyper- and hypoglycemia among

inpatients with diabetes: A national survey of 44 U.S. hospitals. Diabetes Care, 30(2), 367-369.

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

**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*)

**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*)

**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*)  
**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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*)

We quantitatively assessed data element feasibility using the rate of missing for each required EHR data element for measure calculation.

For the EHR data elements used in this measure, we anticipate that there may be some missing data. However, we included only those variables that we expect to be consistently obtained in the target population, available in structured fields, and captured as part of the standard care workflow.

**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*)

Table 7. Frequency of Missing Data by Data Element Required for Measure (Beta Datasets 1 and 2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Beta Dataset 1** | | | **Beta Dataset 2** | | |
| **Missing  Count (#)** | **Encounters (#)** | **Missing Percent (%)** | **Missing  Count (#)** | **Encounters (#)** | **Missing Percent (%)** |
| Admission characteristic: admission date and time | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |
| Antihyperglycemic medication administered: order ID | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |
| Antihyperglycemic medication administered with date and time | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |
| Laboratory test, blood glucose results | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |
| Laboratory test, blood glucose date and time | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |
| Patient characteristic: birth date | 0 | 7,748 | 0.0% | 0 | 5,888 | 0.0% |

**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*)

Among the data elements required for the measure calculation, there were no missing data meaning all encounters had all required data elements, showing that it was feasible to extract the data elements from each test site’s EHR. This means each encounter had an antihyperglycemic medication, and a blood glucose result. Because all patients in the measure denominator received at least one antihyperglycemic medication during their hospitalization, we expect all patients to receive glucose blood tests.