**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: Glycemic Control – Hyperglycemia

**Date of Submission**: 12/5/2013

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| ☐ Cost/resource | ☐ Process |
| ☐ 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, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** 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 (2b2-2b6) 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 20 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). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering 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.  **2b2.** **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.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) 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.   **2b5.** 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.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **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.  **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.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** 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.23*)** | **Measure Tested with Data From:** |
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☐ administrative claims | ☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/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*).

Not applicable

**1.3. What are the dates of the data used in testing**? Admissions from January 3, 2011 – December 31, 2012

**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.26*)** | **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*)   
A total of eight hospitals, comprising seven acute care hospitals and one critical access hospital (CAH), from four states (AZ, FL, MO, and TX) were used to perform the field testing of the measure. FMQAI included a variety of hospital sites within its field testing sites to encompass urban and rural, large and small, teaching and non-teaching, and for-profit and non-profit types. In addition, FMQAI included hospitals with EHR systems with the largest market share, such as Epic and Cerner, as well as a hospital with a home-grown, hybrid version of EHR systems that qualified for Stage I Meaningful Use. Table 1 provides a breakdown of the characteristics of the participating hospitals included in the field testing of the measures.

**Table 1. Field Testing Hospital Characteristics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID** | **Location** | **Size** | **Type** | **Teaching Status** | **Ownership** | **EHR System** |
| 1 | Urban | 852 | Level 1 Trauma | Teaching | Non-profit | Epic |
| 2 | Urban | 522 | Level 1 Trauma | Teaching | Non-profit | Epic |
| 3 | Rural | 25 | Critical Access | Non-Teaching | Non-profit | Cerner |
| 4 | Urban | 312 | Acute Regional | Teaching | For Profit | McKesson |
| 5 | Urban | 493 | Level 1 Trauma | Non-Teaching | For Profit | Cerner |
| 6 | Urban | 359 | Acute Care Community | Teaching | For Profit | Cerner |
| 7 | Urban | 143 | Acute Care Community | Teaching | For Profit | Cerner |
| 8 | Urban | 695 | Level 1 Trauma | Teaching | Non-profit | Epic |

**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*)   
EHR data extracts obtained from the eight participating hospitals in four states (AZ, FL, MO, and TX) were used to conduct the field testing of the measure. The criterion for the field testing sample was to include approximately 5,000 inpatient admissions or at least one year of admissions per hospital. Table 2 shows the demographic characteristics of the sample by hospital.

**Table 2. Demographic Characteristics of the Field-Testing Sample**

| **Demographic** | **Hospital 1** | | **Hospital 2** | | **Hospital 3** | | **Hospital 4** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  | |  | |  | |  |
| Female | 4,531 (55.05%) | | 3,095 (51.51%) | | No data | | 5,563 (50.80%) |
| Male | 3,700 (44.95%) | | 2,914 (48.49%) | | No data | | 5,388 (49.20%) |
| **Age** |  | |  | |  | |  |
| ≥ 65 years | 2,257 (27.42%) | | 571 (9.5%) | | 153 (68.00%) | | 8,615 (78.67%) |
| **Race** |  | |  | |  | |  |
| White/Caucasian | 5,658 (68.74%) | | 2,338 (38.91%) | | 184 (81.78%) | | 10,762 (98.72%) |
| African-American | 1,792 (21.77%) | | 733 (12.20%) | | 17 (7.56%) | | 51 (0.47%) |
| Hispanic | 353 (4.29%) | | 2,502 (41.64%) | | 0 | | 32 (0.30%) |
| Other | 428 (5.20%) | | 436 (7.25%) | | 24 (10.66%) | | 106 (0.97%) |
| **Ethnicity** |  | |  | |  | |  |
| Hispanic | 309 (3.75%) | | 2,318 (38.58%) | | 27 (12.00%) | | 16 (0.15%) |
| Non-Hispanic | 7,921 (96.25%) | | 3,691 (61.42%) | | 198 (88.00%) | | 10,935 (99.85%) |
| **Payor Source** |  | |  | |  | |  |
| Medicare | 3,100 (37.66%) | | 926 (15.41%) | | 185 (82.22%) | | 8,698 (79.43%) |
| Medicaid | 1,682 (20.43%) | | 0 | | 12 (5.33) | | 292 (2.67%) |
| Self-Pay | 951 (11.55%) | | 2,029 (33.77%) | | 6 (2.67%) | | 336 (3.07%) |
| Other | 2,498 (30.35%) | | 3,054 (50.82%) | | 22 (9.78%) | | 1,625 (14.84%) |
| **Demographic** | | **Hospital 5** | | **Hospital 6** | | **Hospital 7** | **Hospital 8** |
| **Gender** | |  | |  | |  |  |
| Female | | 4,320 (49.57%) | | 5,792 (68.59%) | | 4,722 (54.89%) | 3,005 (55.80%) |
| Male | | 4,395 (50.43%) | | 2,652 (31.41%) | | 3,880 (45.11%) | 2,380 (44.20%) |
| **Age** | |  | |  | |  |  |
| ≥ 65 years | | 5,988 (68.71%) | | 3,215 (38.07%) | | 4,380 (50.92%) | 1,102 (20.46%) |
| **Race** | |  | |  | |  |  |
| White/Caucasian | | 7,785 (89.33%) | | 7,933 (93.95%) | | 7,592 (88.26%) | 2,499 (44.36%) |
| African-American | | 615 (7.06%) | | 133 (1.58%) | | 798 (9.28%) | 2,814 (49.95%) |
| Hispanic | | 0 | | 0 | | 0 | 0 |
| Other | | 315 (3.61%) | | 378 (4.47%) | | 212 (2.46%) | 321 (5.69%) |
| **Ethnicity** | |  | |  | |  |  |
| Hispanic | | 290 (3.33%) | | 6,445 (76.33%) | | 52 (0.60%) | 237 (4.21%) |
| Non-Hispanic | | 8,425 (96.67%) | | 1,999 (23.67%) | | 8,550 (99.40%) | 5,397 (95.79%) |
| **Payor Source** | |  | |  | |  |  |
| Medicare | | 6,241 (71.61%) | | 3,571 (42.29%) | | 5,229 (60.79%) | 1,777 (33.00%) |
| Medicaid | | 367 (4.21%) | | 1,848 (21.89%) | | 406 (4.72%) | 1,927 (35.78%) |
| Self-Pay | | 492 (5.65%) | | 631 (7.47%) | | 187 (2.17%) | 50 (0.93%) |
| Other | | 1,615 (18.53%) | | 2,394 (28.35%) | | 2,780 (32.32%) | 1,631 (30.29%) |

**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**.

Subsamples were randomly selected from the testing sample described in Section 1.6 for testing the criterion and construct validity of the measure.

Criterion (Content) Validity

The subsample for the criterion validity testing consisted of a random selection of 50 inpatient admissions from each of the 8 participating hospitals (n=400). Three hundred ninety-eight of the 400 cases were reviewed by a trained registered nurse abstractor. Of these cases, 121 (30%) were selected for re-review by another trained registered nurse abstractor.

Construct Validity

The subsample for the construct validity testing consisted of a random selection of 119 inpatient admissions from the overall inpatient admissions received from the 8 participating hospitals. All 119 cases were reviewed by trained physician reviewers. Of these cases, 23 (19%) were selected for re-review by another trained physician reviewer.

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

**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*)

The accuracy/correctness (validity) of data elements for this measure was empirically tested. Therefore, per the NQF Measure Evaluation Criteria and Guidance, separate reliability testing of data elements was not required. Please refer to the Validity Testing section for additional testing information on the scientific acceptability of the measure properties.

In order to assess measure precision in the context of the observed variability across measurement units (hospital facilities), we utilized the approach proposed by Adams (2009). The rationale for this choice of testing was based on the work on the reliability for provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial published by Adams: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.”

The signal-to-noise ratio was calculated as a function of the variance between hospitals (signal) and the variance within a hospital (noise). Reliability was estimated using a beta-binomial model. This approach has two basic assumptions:

1. Each hospital has a true pass rate, p, which varies from hospital to hospital, and
2. The hospital’s score is a binomial random variable conditional on the hospital’s true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual hospital variance), whereas, a reliability of 1.0 implies that all variation is caused by a real difference in performance (across hospitals). In a simulation, Adams showed that differences between providers started to be detectable at reliability of 0.7, and significant differences could be seen at reliability of 0.9. Our rationale was based on Adams’ work, and thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between hospitals.

Citation

Adams, J. L. The reliability of provider profiling: A tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009.

**2a2.3. For each level 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*)  
Reliability tests were conducted across hospitals, and the results are displayed in Table 3.

**Table 3. Reliability Statistics by Hospital**

| **Hospital** | **Denom** | **Measure Rate** | **Reliability Score** |
| --- | --- | --- | --- |
| Hospital 1 | 2,033 | 22.3% | 0.97 |
| Hospital 2 | 0,853 | 27.9% | 0.92 |
| Hospital 3 | 0,074 | 33.0% | 0.42 |
| Hospital 4 | 2,747 | 28.7% | 0.97 |
| Hospital 5 | 2,058 | 24.9% | 0.96 |
| Hospital 6 | 2,215 | 28.5% | 0.96 |
| Hospital 7 | 2,620 | 29.1% | 0.97 |
| Hospital 8 | 1,316 | 26.5% | 0.94 |

**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?*)  
Seven of the eight hospitals have reliability scores that are greater than 0.70, which indicates that the measure would produce reliable scores at the hospital level. Hospital #3 did not achieve an adequate reliability score. However, this is a critical access hospital with only 25 beds, which resulted in a much smaller denominator compared to the other hospitals. The reliability test suggested that the measure rates are reliable across the large and medium-size hospitals.

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

**2b2. VALIDITY TESTING**

**2b2.1. 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*)

**2b2.2. For each level 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)*  
Criterion (Content) Validity

Criterion validity evaluated whether the measure correctly identified all the data elements that were required to calculate the measure rate. In addition, this analytic method quantified the agreement between the extraction of data obtained directly from the participating providers’ EHR systems using the QDM logic applied to the data elements obtained by review and abstraction of the entire EHR as a gold standard. The criterion chart review was performed by registered nurses trained by FMQAI using a standardized chart review electronic abstraction tool developed in Microsoft Access. Nurse abstractors were instructed to verify that the received data elements that were required to calculate the measure rate from the data extraction matched the information contained within the EHR.

Construct Validity

Construct validity addressed whether the measure identified true manifestation of hyperglycemic events as defined by the measure specifications. In addition, the construct validity testing assessed whether additional causes could be attributed to the hyperglycemic event to inform risk factors that would need to be considered for risk stratification purposes. The construct validity chart review was conducted by physician reviewers trained by FMQAI and the University of Florida, using a standardized chart review paper abstraction tool.

Using the above-described sample of charts, the reviewers were asked to assess construct validity using the following quantitative and qualitative questions:

* 1. Did the patient have sustained hyperglycemia (as defined by the measure) at the indicated days? (Yes/No)
  2. Did you identify additional causes that can explain the occurrence of hyperglycemia? (Yes/No)

If so, do you suggest specific variables that should be included in risk adjustment algorithms if hospitals were to be compared?

Face Validity

FMQAI’s Technical Expert Panel (TEP) evaluated the face validity of the measure and measure score after field testing was completed. The names and organizations of TEP members are listed in Table 4. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®). TEP members were specifically asked whether “the performance score from the measure as specified represents an accurate reflection of quality of care.” They responded by indicating their level of agreement with the statement on a 5-point Likert scale (1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree).

**Table 4. TEP Members**

| **Name** | **Organization** |
| --- | --- |
| **Dale W. Bratzler**  DO, MPH | Professor and Associate Dean, College of Public Health, University of Oklahoma Health Sciences Center |
| **Mary Brennan-Taylor** | Adjunct Research Instructor of Family Medicine, School of Medicine and Biomedical Sciences, University of Buffalo  Representing: TEP as Patient Representative |
| **Frank E. Briggs III**  PharmD, MPH | Vice President, Quality and Patient Safety, West Virginia University Healthcare  Representing: American Society of Health-System Pharmacists |
| **Daniel Castillo**  MD, MBA | Medical Director, Healthcare Quality Evaluation, The Joint Commission |
| **Joan Ching**  RN, MN, CPHQ | Administrative Director, Hospital Quality & Safety, Virginia Mason Medical Center |
| **Edward S. Eisenberg**  MD, FACP | Senior Vice President, Performance Measurement and Strategic Alliances, Pharmacy Quality Alliance |
| **Floyd Eisenberg**  MD, MPH, FACP | President, iParsimony, LLC |
| **Marybeth Farquhar**  PhD, MSN, RN | Vice President of Research & Measurement, URAC |
| **Frank Federico**  BS, RPh | Executive Director for Strategic Partners, Institute for Healthcare Improvement |
| **Robert Feroli**  PharmD, FASHP | Medication Safety Officer, Johns Hopkins Hospital |
| **Tejal Gandhi**  MD, MPH | President, National Patient Safety Foundation; Board-certified Internist and Associate Professor of Medicine, Harvard Medical School  Representing: American Hospital Association |
| **P. Michael Ho**  MD, PhD, FACC | Staff Cardiologist, VA Eastern Colorado Health Care System; Associate Professor of Medicine, University of Colorado Denver  Representing: American College of Cardiology |
| **Mark L. Holtsman**  PharmD | Co-Director, Inpatient Pain Service and Pain Management Service Pharmacist, UC Davis Medical Center; Clinical Professor of Anesthesiology and Pain Medicine, UC Davis School of Medicine  Representing: American Academy of Pain Medicine |
| **Clifford Ko**  MD, MS, MSHS, FACS | Director, ACS Division of Research and Optimal Patient Care; Director, ACS NSQIP; Professor of Surgery and Health Services, UCLA Schools of Medicine and Public Health  Representing: American College of Surgeons |
| **Janet Maurer**  MD, MBA, FCCP | Operations Medical Director, National Imaging Associates, Health Dialog; Clinical Professor of Medicine, University of Arizona, College of Medicine, Phoenix Campus; Staff Physician, St. Joseph’s Medical Center  Representing: American College of Chest Physicians |
| **Michael N. Neuss**  MD | Chief Medical Officer, Vanderbilt-Ingram Cancer Center; Professor of Medicine, Vanderbilt School of Medicine  Representing: American Society of Clinical Oncology |
| **N. Lee Rucker**  MSPH | Senior Advisor, National Council on Patient Information and Education |
| **Edward Septimus**  MD, FACP, FIDSA, FSHEA | Medical Director, Infection Prevention and Epidemiology Clinical Service Group, HCA Healthcare System; Clinical Professor of Internal Medicine, Texas A & M University  Representing: Infectious Diseases Society of America |
| **Nathan Spell**  MD, FACP | Chief Quality Officer, Emory University Hospital; Associate Professor of Medicine, Emory University School of Medicine  Representing: American College of Physicians |
| **Stephen J. Traub**  MD, FACEP | Assistant Professor in Emergency Medicine and Chair, Department of Emergency Medicine, Mayo Clinic  Representing: American College of Emergency Physicians |
| **Darren M. Triller**  PharmD | Senior Director, Quality Improvement, IPRO QIO |

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Criterion (Content) Validity

Criterion validity was summarized as the percent of data elements that were correctly identified by the electronically extracted measure, according to the chart review. As noted above, inter-rater reliability, or proportion of agreement concerning the accuracy of the data elements received from the data extract, was calculated using the 398 charts that were reviewed by the nurse abstractors. The proportion of agreement on each data element is shown in Table 5 by hospital.

Overall, 98.17% of the data elements found in the medical record correctly matched the EHR data extract received from the participating hospitals. The data element with the lowest criterion validity score was the “Location: ICU DateTime” at 83.41%. Analysis of the “Location: ICU DateTime” data element showed a wide variation between hospitals, ranging from 60.00% (Hospital 8) to 100% (Hospitals 3 and 4) match. Chart review findings, when compared to the data extract received from the participating hospitals, revealed inconsistent documentation of dates and times when a patient was transferred from one unit to another. Further investigation of the data received from the hospitals with the lowest scores revealed that the time period for the data received for the “Location DateTime” data element was prior to the implementation of the hospital’s EHR system and derived from the Universal Billing (UB) charge level, which was only recorded once each day at midnight. The UB charge level information did not necessarily align with the clinical documentation within the record.

**Table 5. Results of Criterion Validity Testing**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Hospital 1** | **Hospital 2** | **Hospital 3** | **Hospital 4** | **Hospital 5** | **Hospital 6** | **Hospital 7** | **Hospital 8** | **Average** |
| Date\_of\_Birth | N/A | 100% | 91.84% | 100% | 100% | 100% | 100% | 100% | 98.83% |
| Admission\_DateTime | 93.88% | 100% | N/A | 100% | 100% | 100% | 100% | 100% | 99.13% |
| Discharge\_DateTime | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Diagnosis\_Code – DM | 97.96% | 100% | 100% | 98.00% | 100% | 100% | 100% | 100% | 99.50% |
| Diagnosis\_Code - DKA | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Diagnosis\_Code - HHS | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Medication\_Administered\_  DateTime – Anti-diabetic | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Lab\_Test - BG > 200 | 100% | 100% | 100% | 100% | 98.00% | 100% | 100% | 98.00% | 99.50% |
| Lab\_Collection\_DateTime –BG > 200 | 100% | 100% | 100% | 100% | 98.00% | 96.00% | 98.00% | 98.00% | 98.75% |
| Lab\_Test – First BG > 400 | 100% | 100% | 100% | 100% | 98.00% | 100% | 100% | 100% | 99.75% |
| Procedure\_Code – CT Surg. | 97.96 | 100% | N/A | 100% | 100% | 100% | 100% | 100% | 99.75% |
| Location: ICU DateTime | 67.35% | 98.00% | N/A | 100% | 84.00% | 70.00% | 88.00% | 60.00% | 81.05% |
| Total Proportion Agreement | 96.10% | 99.83% | 99.09% | 99.83% | 98.17% | 97.17% | 98.83% | 96.33% | 98.17% |

One hundred and twenty-one cases were re-reviewed by a different nurse abstractor to check the reliability of the abstraction. Of these hyperglycemic cases reviewed by both nurse abstractors, the inter-rater reliability (proportion of agreement) score was 97.8%. Since this value was above the 95% minimum acceptable score, no further re-reviews were required. All discrepancies found were analyzed and discussed by both abstractors as an additional training tool to aid in quality improvement for future abstractions.

Construct Validity

Construct validity was summarized as the percent of cases detected by the measure that were true cases of manifest hyperglycemic events, and the test results showed that 100% of the 119 cases reviewed had a true hyperglycemic manifestation confirmed by the physician reviewers. For the hyperglycemic cases reviewed by both physician reviewers, there was 100% agreement between reviewers.

Face Validity

Eighteen of the 21 TEP members participated in the face validity evaluation for the measure. The results of the TEP rating of face validity on a scale of 1 to 5 are presented in Table 6.

**Table 6. Results of Face Validity Evaluation**

|  |  |
| --- | --- |
| **Rating** | **Number of TEP Members** |
| 5 (Strongly Agree) | 4 |
| 4 (Agree) | 11 |
| 3 (Neutral) | 2 |
| 2 (Disagree) | 1 |
| 1 (Strongly Disagree) | 0 |

The majority of TEP members agreed that the measure was valid (83.3%), and both the mean and median score were 4, indicating overall the TEP considered that the measure was valid as specified.

**2b2.4. 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?*)  
The results from content validity testing demonstrated that the data elements in the data extract generated from the EHR systems are accurate when compared to manual abstraction of the full medical record from the EHR. The results from the construct validity indicate that the measure accurately captures inpatient hyperglycemia as defined by the specifications. Concerning the validity of the “Location: ICU DateTime” data element, the results demonstrate wide variability across hospitals resulting from inconsistent documentation of unit transfer times by the transferring and receiving units. In addition, the time inconsistencies between the data extracted from EHR systems and the recorded times in the medical record by the clinical staff varied across clinical specialty (physician, nurse, social worker, physical therapy, etc.). However, we found that the receiving unit times were more likely to be accurate. Lastly, 15 of the 18 TEP members either responded “agree” or “strongly agree” that the measure, as specified, exhibited face validity.

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

**2b3. EXCLUSIONS ANALYSIS**

**NA** ☐ **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. 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*)  
The following 3 exclusions apply to the measure:

1. Admissions with diagnosis of diabetic ketoacidosis (DKA) or hyperglycemic hyperosmolar syndrome (HHS) are excluded.
2. Admissions without any hospital days included in the analysis are excluded.
3. Admissions with lengths of stay greater than 120 days are excluded.

The TEP work group suggested excluding admissions with DKA or HHS. Guidelines suggest that patients with DKA require more careful downward-titration of glucose values because of a greater risk of hypoglycemia, rebound ketosis derived by counter-regulatory hormones, and the fact that rapid correction of hyperosmolarity may shift water to the hyperosmolar intracellular space and induce cerebral edema.

To examine the effect of these exclusions, the number affected by exclusion was first examined and the measure rates with and without each exclusion were calculated and compared.

**2b3.2. 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*)

**Table 7. Exclusion of Qualified Denominator Admissions with Diagnosis of DKA or HHS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Hospital** | **Number Excluded (%)** | **Measure Rate With Exclusion** | **95% Confidence Interval (CI)** | **Measure Rate Without Exclusion** | **95% CI** |
| Hospital 1 | 39 (1.88%) | 22.3% | 20.7%-23.9% | 23.2% | 21.6%-24.8% |
| Hospital 2 | 27 (3.07%) | 27.9% | 25.5%-30.4% | 29.3% | 26.8%-31.7% |
| Hospital 3 | 0 | 33.0% | 24.5%-41.4% | 33.0% | 24.5%-41.4% |
| Hospital 4 | 32 (1.15%) | 28.7% | 27.4%-30.1% | 29.0% | 27.6%-30.4% |
| Hospital 5 | 25 (1.20%) | 24.9% | 23.4%-26.5% | 25.3% | 23.8%-26.9% |
| Hospital 6 | 60 (2.64%) | 28.5% | 27.0%-30.1% | 29.1% | 27.6%-30.6% |
| Hospital 7 | 11 (0.42%) | 29.1% | 27.7%-30.5% | 29.2% | 27.7%-30.6% |
| Hospital 8 | 35 (2.59%) | 26.5% | 24.5%-28.5% | 27.8% | 25.8%-29.7% |

**Table 8. Exclusion of Qualified Denominator Admissions without Any Hospital Days Left for Analysis**

|  |  |
| --- | --- |
| **Hospital** | **Number Excluded (%)** |
| Hospital 1 | 614 (23.20%) |
| Hospital 2 | 480 (35.29%) |
| Hospital 3 | 24 (24.49%) |
| Hospital 4 | 684 (19.75%) |
| Hospital 5 | 572 (21.54%) |
| Hospital 6 | 594 (20.70%) |
| Hospital 7 | 639 (19.54%) |
| Hospital 8 | 477 (26.09%) |

The admissions listed in Table 8 did not contribute any measureable days in the measure, and therefore, the measure rates were not impacted by this exclusion.

**Table 9. Exclusion of Qualified Denominator Admissions with Lengths of Stay Greater than 120 Days**

|  |  |  |  |
| --- | --- | --- | --- |
| **Hospital** | **Number Excluded (%)** | **Measure Rate with Exclusion** | **Measure Rate without Exclusion** |
| Hospital 5 | 1 (0.05%) | 25.35% | 25.34% |

The other 7 hospitals did not have any admissions with lengths of stay greater than 120 days that were excluded from the measure denominator.

**2b3.3. 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*)  
Although there is no statistical difference between measure rates with or without the exclusion of DKA or HHS patients, after careful review of treatment guidelines for DKA and HHS, the TEP work group agreed that goals for the achievement of normoglycemia apply, but that downward-titration of glucose levels needs to be slower than hyperglycemic patients without DKA/HHS. Since there are distinct differences in treatment goals and the number of patients affected by this exclusion is a relatively small proportion of the denominator, the TEP work group recommended the exclusion of admissions with DKA or HHS entirely.

A total of 4,084 admissions in the field testing data sample were not counted in the measure, because these admissions did not contribute any measureable days after the denominator logic was applied. The denominator logic of the measure was specifically built to allow for a grace period for clinicians to implement interventions for managing blood glucose level and for patients to react to the interventions. When the logic is applied, it can completely exclude admissions with a very short length of stay (≤ 3 days) recognizing that it could be challenging for the clinicians to effectively manage hyperglycemia when the patient’s length of stay was less than three days. On average across the field testing hospitals 418 measurable days were excluded due to a very short length of stay.

There was only one admission that was excluded from the measure denominator due to a length of stay greater than 120 days. The exclusion was maintained to align the proposed measure with other measures in the CMS Hospital Inpatient Quality Reporting Program.

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

**2b4. 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*** [***2b5***](#section2b5)***.***

**2b4.1. 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** 3 **risk categories**

☐ **Other,** Click here to enter description

**2b4.2. If an outcome or resource use 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**.   
Not applicable

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)

Potential patient factors for measure stratification were identified through a literature review, suggestions from clinical experts, and the chart review during the formative testing phase that included an open-ended question about possible risk factors presented in the hyperglycemia cases reviewed. These potential risk factors were then discussed with the TEP.

**2b4.4. What were the statistical results of the analyses used to select risk factors?**Not applicable

**2b4.5. 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*)  
The literature review identified a large number of risk factors for hyperglycemia; however, there was clear evidence indicating that many of the risk factors can be addressed with appropriate titration of anti-hyperglycemic treatment. Thus, even though the patient case mix may result in different rates of hyperglycemia if unaddressed, appropriate therapeutic adjustments are expected to result in similar rates. Exceptions that were determined during chart review of numerator cases included hyperglycemia after administration of high doses of steroids. Clinical experts agreed that individual patient responses to steroids vary, resulting in different and difficult to predict insulin needs. Thus, certain cases of hyperglycemia following administration of high doses of steroids were considered unpreventable, as insulin doses had been increased following current recommendations and patients’ subsequent glucose values were unpredictably higher. Importantly, clinical experts suggested that at least a proportion of such cases are preventable because minimal increases in insulin should be made and proper insulin titration should occur immediately once hyperglycemia is noted. Thus, because high-dose steroid use is common in Medicare recipients (e.g., in those with COPD exacerbations) and because at least a portion of hyperglycemic events are preventable, the TEP made a recommendation to risk stratify the measure. Following the same reasoning, the TEP recommended stratification by intensive versus acute care units as well as surgical versus medical patients. In both instances (post-surgery and in intensive care) patients are expected to be at higher risk for stress-induced hyperglycemia, which presents similar challenges as the management of steroid-induced hyperglycemia.

To investigate the effect of risk factors, a stratified analysis was conducted, and the stratified rates were compared for each hospital.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**Not applicable

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not applicable

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Not applicable

**2b4.9. Results of Risk Stratification Analysis**:

To measure a hospital’s performance with managing hyperglycemia, the metric examined the average percentage of hospital days that at-risk patients were hyperglycemic. The numerator was the sum of the percentage of hospital days where the patient was hyperglycemic for each admission. The denominator was the number of qualifying admissions. The average measure rate indicates the mean measure performance across the hospitals.

The measure rates are stratified by setting (ICU vs. non-ICU), patient type (medical vs. surgical), and steroid daily cumulative dose (≤10 mg vs. >10 mg), and the results are displayed in the following tables.

**Table 10. Denominators, Numerators, and Measure Rates Stratified by Settings[[1]](#footnote-2)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID** | **ICU Days** | | | | **Non-ICU Days** | | | |
| **Denom** | **Num** | **Measure Rate** | **95% CI** | **Denom** | **Num** | **Measure Rate** | **95% CI** |
| **1** | 755 | 116.7 | 15.5% | 13.0%-18.2% | 1781 | 407.6 | 22.9% | 21.0%-24.9% |
| **2** | 207 | 040.2 | 19.4% | 14.4%-25.5% | 0703 | 201.5 | 28.7% | 25.4%-32.2% |
| **3** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **4** | 643 | 163.1 | 25.4% | 22.1%-28.9% | 2589 | 732.1 | 28.3% | 26.6%-30.1% |
| **5** | 423 | 096.0 | 22.7% | 18.9%-27.0% | 1844 | 464.9 | 25.2% | 23.3%-27.3% |
| **6** | 450 | 119.7 | 26.6% | 22.7%-30.9% | 2051 | 582.0 | 28.4% | 26.5%-30.4% |
| **7** | 598 | 156.0 | 26.1% | 22.7%-29.8% | 2406 | 710.6 | 29.5% | 27.7%-31.4% |
| **8** | 327 | 068.5 | 21.0% | 16.8%-25.8% | 1174 | 324.0 | 27.6% | 25.1%-30.3% |

N/A = Not applicable

**Table 11. Denominators, Numerators, and Measure Rates Stratified by Admission Type1**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID** | **Surgical Admissions** | | | | **Medical Admissions** | | | |
| **Denom** | **Num** | **Measure Rate** | **95% CI** | **Denom** | **Num** | **Measure Rate** | **95% CI** |
| **1** | 0905 | 151.6 | 16.8% | 14.4%-19.4% | 1128 | 301.5 | 26.7% | 24.2%-29.4% |
| **2** | 0272 | 72.4 | 26.6% | 21.6%-32.3% | 0581 | 165.9 | 28.6% | 25.0%-32.4% |
| **3** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **4** | 1042 | 240.3 | 23.1% | 20.6%-25.7% | 1705 | 549.0 | 32.2% | 30.0%-34.5% |
| **5** | 0591 | 139.8 | 23.6% | 20.4%-27.3% | 1467 | 373.4 | 25.5% | 23.3%-27.8% |
| **6** | 0766 | 187.7 | 24.5% | 21.6%-27.7% | 1449 | 444.4 | 30.7% | 28.3%-33.1% |
| **7** | 1243 | 324.0 | 26.1% | 23.7%-28.6% | 1381 | 438.6 | 31.8% | 29.4%-34.3% |
| **8** | 0402 | 077.3 | 19.2% | 15.6%-23.4% | 0914 | 271.5 | 29.7% | 26.8%-32.8% |

N/A = Not applicable

**Table 12. Denominators, Numerators, and Measure Rates Stratified by Steroid Dose[[2]](#footnote-3)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID** | **Daily Cumulative Dose**  **≤10 mg** | | | | **Daily Cumulative Dose**  **>10 mg** | | | |
| **Denom** | **Num** | **Measure Rate** | **95% CI** | **Denom** | **Num** | **Measure Rate** | **95% CI** |
| **1** | 1864 | 379.4 | 20.4% | 18.8%-22.0% | 418 | 145.5 | 34.8% | 30.8%-38.8% |
| **2** | 812 | 217.4 | 26.8% | 24.4%-29.2% | 82 | 32.5 | 39.6% | 30.5%-48.7% |
| **3** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **4** | 2453 | 619.3 | 25.3% | 23.9%-26.6% | 554 | 266.8 | 48.2% | 44.7%-51.6% |
| **5** | 1843 | 404.9 | 22.0% | 20.4%-23.6% | 340 | 140.3 | 41.3% | 36.8%-45.7% |
| **6** | 1985 | 497.2 | 25.1% | 23.5%-26.6% | 444 | 207.5 | 46.7% | 42.8%-50.6% |
| **7** | 2374 | 624.9 | 26.3% | 24.9%-27.7% | 392 | 198.9 | 50.7% | 46.6%-54.9% |
| **8** | 1146 | 251.8 | 22.0% | 19.9%-24.0% | 283 | 132.6 | 46.8% | 42.0%-51.7% |

**2b4.10. 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*)  
Most but not all field testing hospitals exhibited increasing rates of hyperglycemia with increasing doses of steroids. The variation in the degree of these increases suggests that hospitals may have different practices to address hyperglycemic steroid effects. However, the fact that no hospital was able to compensate in its management of hyperglycemia completely for steroid effects emphasizes the need for stratification.

Interestingly, non-ICU days showed consistently higher hyperglycemia rates than ICU days even though the risk for stress-induced hyperglycemia could potentially be larger in ICU environments. Most likely, the greater clinical emphasis on blood glucose management in ICUs along with the ability to use insulin infusions in this setting may have contributed to these results. The various root causes notwithstanding, the analysis demonstrates a fairly homogenous difference between ICU and non-ICU days across all hospitals, which was adequately addressed with stratification.

Similarly, all field testing hospitals showed larger hyperglycemia rates in medical rather than surgical patients. Again, greater emphasis in clinical guidelines as well as more comprehensive management post-surgically especially in the ICU may have contributed to these results. Because surgery did not appear to be a significant risk factor and other explanations appear to lie more in environmental or manageable factors, stratification may not be necessary.

As suggested earlier, other risk factors such as infections may increase the risk for hyperglycemia, but evidence suggests that such increases can be compensated with respective changes in anti-hyperglycemic management. Because the availability of insulin infusions, which are superior to other anti-hyperglycemic regimen in managing hyperglycemia, is typically restricted to intensive care units due to larger monitoring and thus staffing needs, we conclude that patient location should be used for stratification. Likewise, because of a certain unpredictability of insulin needs after exposure to large doses of corticosteroids, we suggest stratification.

**2b4.11.** **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*)

Not applicable

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

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. 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)*

To identify statistically significant differences in performance, chi-square tests of homogeneity were conducted to detect differences at the hospital level. If significant differences were identified, pairwise comparisons were then used to identify differences between specific hospitals.

For the measure rates stratified by steroid dose, a comparison of means at the hospital level was performed. Confidence intervals (95%) were calculated around point estimates for each hospital and then compared to the average performance rate across the hospitals. If the confidence intervals did not overlap with the average measure rate, the difference was considered statistically significant.

**2b5.2. 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 results are displayed in Tables 13, 14, and 15 by stratification categories, and Table 16 showed the non-stratified rate by hospitals. The mean measure rate indicated the average measure performance across hospitals.

**Table 13. Meaningful Differences in Measure Rates Stratified by Settings[[3]](#footnote-4)**

|  |  |  |
| --- | --- | --- |
| **Hospital ID** | **ICU Days Measure Rate** | **Non-ICU Days Measure Rate** |
| **1** | 15.5% | 22.9% |
| **2** | 19.4% | 28.7% |
| **3** | N/A | N/A |
| **4** | 25.4% | 28.3% |
| **5** | 22.7% | 25.2% |
| **6** | 26.6% | 28.4% |
| **7** | 26.1% | 29.5% |
| **8** | 21.0% | 27.6% |
| **Mean** | **22.4%** | **27.2%** |

N/A = Not applicable

ICU Days: The measure rate for Hospital 1 was statistically significantly lower than 22.7% (Hospitals 4, 5, 6, and 7, p-value ≤ 0.014).

Non-ICU Days: The rate for Hospital 1 was statistically significantly lower than 27.6% (Hospitals 2, 4, 6, 7, and 8, p-value ≤ 0.013). The rate for Hospital 7 was statistically significantly higher than Hospital 5 (p-value=0.029).

**Table 14. Meaningful Differences in Measure Rates Stratified by Admission Type[[4]](#footnote-5)**

|  |  |  |
| --- | --- | --- |
| **Hospital ID** | **Surgical Admissions Measure Rate** | **Medical Admissions Measure Rate** |
| **1** | 16.8% | 26.7% |
| **2** | 26.6% | 28.6% |
| **3** | N/A | N/A |
| **4** | 23.1% | 32.2% |
| **5** | 23.6% | 25.5% |
| **6** | 24.5% | 30.7% |
| **7** | 26.1% | 31.8% |
| **8** | 19.2% | 29.7% |
| **Mean** | **22.8%** | **29.3%** |

N/A = Not applicable

Surgical Admissions: The rate for Hospital 1 was statistically significantly lower than all other hospitals except Hospital 8 (p-value ≤0.017). The rate for Hospital 7 was significantly higher than Hospital 8 (p-value=0.005).

Medical Admissions: The rate for Hospital 5 was statistically significantly lower than Hospitals 4, 6, and 7 (p-value ≤ 0.005). The rate for Hospital 1 was statistically significantly lower than Hospitals 4 and 7 (p-value ≤ 0.024).

**Table 15. Meaningful Differences in Measure Rate Stratified by Steroid Dose[[5]](#footnote-6)**

|  |  |  |
| --- | --- | --- |
| **Hospital ID** | **Daily Cumulative Dose ≤10 mg Measure Rate** | **Daily Cumulative Dose >10 mg Measure Rate** |
| **1** | 20.4% | 34.8% |
| **2** | 26.8% | 39.6% |
| **3** | N/A | N/A |
| **4** | 25.3% | 48.2% |
| **5** | 22.0% | 41.3% |
| **6** | 25.1% | 46.7% |
| **7** | 26.3% | 50.7% |
| **8** | 22.0% | 46.8% |
| **Mean** | **24.0%** | **44.0%** |

N/A = Not applicable

Daily Cumulative Dose **≤**10 mg: The rate for Hospital 1 was significantly lower than Hospitals 2, 4, 6, and 7 (p-value ≤0.007). The rate for Hospital 7 was significantly higher than both Hospitals 5 and 8 (p-value ≤0.011). The rate for Hospital 2 was also significantly higher than Hospital 5 (p-value= 0.020).

Daily Cumulative Dose >10 mg: The rate for Hospital 1was statistically significantly lower Hospitals 4, 6, 7, and 8 (p-value ≤0.004). The rate for Hospital 5 was significantly lower than Hospital 7 (p-value=0.037)

**Table 16. Meaningful Differences in Non-Stratified Measure Rates**

|  |  |  |
| --- | --- | --- |
| **Hospital ID** | **Admissions** | |
| **Measure Rate** | **95% CI** |
| **1** | 22.3% | 20.5%-24.2% |
| **2** | 27.9% | 25.0%-31.1% |
| **3** | 33.0% | 23.1%-44.9% |
| **4** | 28.7% | 27.1%-30.5% |
| **5** | 24.9% | 23.1%-26.9% |
| **6** | 28.5% | 26.7%-30.5% |
| **7** | 29.1% | 27.4%-30.8% |
| **8** | 26.5% | 24.2%-29.0% |
| **Mean** | **27.6%** | **N/A** |

N/A = Not applicable

For the non-stratified measure rate, Hospital 1 was statistically significantly lower than 26.5% (Hospitals 2, 4, 6, 7, and 8, p-value ≤0.025). Hospitals 4 and 6 were significantly higher than Hospital 5 (p-value=0.0081 and 0.0269, respectively).

**2b5.3. 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?*)  
The results demonstrated that statistically significant differences can be detected between hospitals. The variations in performance among the hospitals suggested meaningful differences in the quality of care provided between the lowest and highest performing hospital and indicated that there is ample room for improvement.

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

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

**2b6.2. 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*)  
Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of 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*)  
Not applicable

1. Hospital #3 is a critical access hospital. It does not have ICU or surgical patients, and therefore, the measure rate cannot be stratified. [↑](#footnote-ref-2)
2. The information related to steroid dosing was not available for Hospital #3, and therefore, the measure rate stratified by daily cumulative steroid dose was not calculated for this hospital. [↑](#footnote-ref-3)
3. Hospital #3 is critical access hospital. It does not have ICU or surgical patients, and therefore, the measure rate cannot be stratified. [↑](#footnote-ref-4)
4. 4 Hospital #3 is critical access hospital. It does not have ICU or surgical patients, and therefore, the measure rate cannot be stratified. [↑](#footnote-ref-5)
5. 5 The information related to steroid dosing was not available for Hospital #3, and therefore, the measure rate stratified by daily cumulative steroid dose was not calculated for this hospital. [↑](#footnote-ref-6)