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

**Measure Number** (*if previously endorsed*)**:** 0640

**Measure Title**:  Hours of physical restraint use

**Date of Submission**: 12/20/2018

**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 (2 a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

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

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

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

**1.1. 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*). Not applicable

**1.3. What are the dates of the data used in testing**? 4/1/2007 – 7/1/2007

**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*)

Description of the population characteristics

This measure has been in national use since the 4th quarter of 2008. Demographics of organizations collecting and reporting data on these measures are as follows:

487 Health care organizations representing various types, locations and sizes:

408 Free-Standing Psychiatric Hospitals, 79 Acute-Care Hospitals with Psychiatric Units

103 For Profit, 120 Not for Profit, 184 Government

103 >=300 beds; 217 100-299 beds; 67 <100 beds

States represented in this data collection effort include: AK, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, PR, RI, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY

27 performance measurement systems are used for data transmission to The Joint Commission.

Description of sampling method

Ten hospitals were randomly sampled from the 487 hospitals in the population, using a stratified sampling methodology to represent the three bed size and three ownership categories.

**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*)

Patients were randomly sampled from each of the ten hospitals in the sample, using a stratified sampling methodology so that measure numerator and denominator cases identified in the original abstraction were represented in the sample and an equal number of cases were sampled for each hospital. There were 191 patients sampled in all.

**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 validity and exclusion analysis was done using 2017 data (from patients with a discharge in 2017) from all hospitals submitting the HBIPS measures to The Joint Commission.

**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.

Not applicable, not required at the time this testing was done.

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

**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*)

All sampled cases were re-abstracted by trained Joint Commission staff. Re-abstracted data are compared with originally abstracted data on a data element by data element basis. The test used were the calculated agreement rates for individual data elements that are used to compute measure rates for the measure.

**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*)

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Elements** | **Total Numerator** | **Total Denominator** | **Agreement Rate** |
| **Numerator Data Elements** |  |  |  |
| Event Date | 11 | 11 | 100.0% |
| Event Type | 11 | 11 | 100.0% |
| Minutes of Physical Restraint | 11 | 11 | 100.0% |

The above data elements were assessed for reliability: event date, event type and minutes of physical restraint. There was a 100% match for the calculated agreement rate for these data elements which are used to compute measure rates for the measure.

Additionally, re-abstraction data analysis containing the health care organization’s Category

Assignment Agreement Rate (CAAR) which represents assignment to the numerator or denominator was performed on these same data from the sample hospitals resulting in an agreement rate of 100%.

**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?*)

A perfect agreement rate between originally abstracted data and re-abstracted data equals 100%, and an agreement rate below 75% is considered failing. These agreement rates are considered to be well within acceptable levels.

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

**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)*

At the time this measure was originally tested measure validity was assessed via survey and focus groups of hospitals participating in the pilot test. All measure specifications, including population identification, numerator and denominator statements, and data elements and their definitions were found to be understandable, retrievable, and relevant.

Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically, continually reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure. And finally, the crosswalk from ICD-9-CM diagnosis codes to ICD-10-CM diagnosis codes has been completed and reviewed by the Technical Advisory Panel for face validity. The panel has determined that the intent of the measure has not changed as a result of the conversion. The crosswalk will also be posted in the next version of the specifications manual for public comment during 2013, and results of feedback will be reviewed and incorporated into the crosswalk where indicated.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)   
Tests for correlations between HBIPS-2 and the remaining HBIPS measures (HBIPS-1, HBIPS-3, HBIPS-5) are -0.00313(p=0.9328), 0.21596 (p<0.001), and -0.04720 (p=0.2068), respectively. This indicates that there are no statistically significant correlations between HBIPS-2 and HBIPS-1 and HBIPS-5. There is a slight positive correlation between HBIPS-2 and HBIPS-3. Employing a longitudinal Poisson regression model of hours of restraint with total patient hours as the offset term and the hospital as a random effect yields a significant improvement of rates over time (p<0.0001).

Queries submitted via the automated feedback system have decreased significantly for the HBIPS measure set in the past 3 years. (522 in 2016, 288 in 2017, 187 for 2018 YTD). There have been no major issues with the data elements for this measure. A few updates were made to the Guidelines for Abstraction of the data element Minutes of Physical Restraint to provide clarification for abstracting time in restraints. Notes for Abstraction were additionally updated to clarify the priority for tracking time in restraint vs. seclusion when a patient is placed in restraint and seclusion at the same time.

Analysis of feedback obtained via our automated feedback system reveals only a few submissions regarding specifications for this measure over the past three years. Predominant themes of these submissions involved questions regarding clarification of the use of manual holds with respect to controlling self-harm behaviors for children and during the administration of medication. All manual holds are included as a form of physical restraint. The definition of physical restraint and examples of physical restraint for this measure were taken verbatim from 42 CFR Part 482, Medicare and Medicaid Programs; Hospital Conditions of Participation: Patient’s Rights. Based on feedback from the forensic hospitals, an additional exclusion was added to the measure specifications excluding patients for restraint uses that are forensic or correctional restrictions applied and used by designated hospital security personnel for the purpose of transporting the patient to court off the locked unit.

Face validity was tested by a total of 40 hospitals during May and June 2006. Measure information was sent to the test hospitals for review. In addition, three site visits with focus interviews were conducted. One site visit had a total of nine state hospitals represented. Criterion validity was evaluated during the focus group interviews conducted during the reliability site visits as well as through an online survey that all pilot hospitals were invited to complete.

The measure information form and the data dictionary were evaluated for face validity. The following parts of the measure information form were evaluated: numerator statement, numerator inclusions, numerator exclusions, denominator statement, denominator inclusions, denominator exclusions and an overall understanding of the measure information form. Each area was scored utilizing a five-point likert scale. For each data element, the hospitals were asked to comment on the clarity and understanding of the abstraction guidelines and data definitions. And finally, the data dictionary was reviewed for overall understanding, usefulness and overall. Qualitative analysis was performed on measure feedback received during the focus group interviews and from the online surveys.

A total of 36 hospitals completed the face validity evaluation and rated the overall understanding of the measure as follows: very good n=13, good n=16, average n=6, poor n=1 and very poor n=0. Modifications to improve the understanding and clarity of the measure specifications were made prior to pilot testing based on feedback received from the hospitals during the face validity evaluation. Analysis of the focus group discussions and the online survey revealed a majority of the pilot hospitals recommended moving the measure forward in the final measure set with suggested modifications. Since that time continual feedback from customers does not indicate a change in their perception of the measure. Also, this measure has been evaluated for validity and adopted for use in a national reimbursement program (CMS).

**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?*)

The positive correlation between HBIPS-2 and HBIPS-3 validates the use of these 2 measures for evaluating quality of care in the behavioral health setting.

The measure has considerable face validity which has been improved over time.

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

**2b2. EXCLUSIONS ANALYSIS**

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

Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this measure set.

This denominator exclusion was analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, Inclusion of leave days in the denominator population would artificially lower the measure rate, and would not be a true representation of the hospital’s actual practice. Because of the role of this measure in the current Joint Commission accreditation process this is especially troubling to end users. This concern is the basis for a number of the non-evidence-based exclusions to these measures. The following measure exclusion that was not derived directly from the evidence is as follows:

* Patients on leave

It is important to note that leave days are typically granted in the public hospital setting and very rarely in the private hospital setting.

**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*)  
Identification of Meaningful Differences in Performance (Measure evaluation criterion 2b5)

N= 278 based on a sample from 17 public hospitals in 2011

* Total leave days =12%

Rationale for exclusion:

* Total leave days

**Rationale:** Time in restraints is calculated based on psychiatric inpatient days. Patient leave days are not part of the calculation of inpatient days so these days are not counted.

**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*)

The rationale indicates that patient leave days should not be counted in the calculation for inpatient days.

The incidence of this exclusion is frequent enough to continue to include in the measure specifications.

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

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

Not applicable

**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

Not applicable

**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.**

Not applicable

**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**.

Not applicable

**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?

Not applicable

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

Not applicable

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

Not applicable

**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.**

Not applicable

**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)

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

**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*)

Not applicable

**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*)

Not applicable

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

**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)*

The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organization’s data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization’s performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the “direction of improvement” of the measure.

There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO’s performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCO’s rating. The estimate of the organization’s true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

**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*)  
HBIPS-2 Distribution of Measure Results

2018 2nd Quarter Data:

Scores on this measure: N=720, Mean 0.3749%, SD 1.3965

10th Percentile= 0%

25th Percentile= 0.009%

50th Percentile= 0.057%

75th Percentile= 0.194%

90th Percentile= 1.560%

575 (79.9%) Favorable – results statistically significantly higher than the national rate

76 (10.6%) Neutral – results not significantly different from target range

69 (9.6%) Unfavorable - results statistically significantly lower than the national rate

**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?*)

Employing a longitudinal Poisson regression model of hours of restraint with total patient hours as the offset term and the hospital as a random effect yields a significant improvement of rates over time (p<0.0001). Although there were improvements over time, measure results continue to demonstrate a gap in care. This measure is important to continue improvement in decreasing the rates of patient restraint.

A practically meaningful number of hospitals were identified with substandard performance for this measure, with performance significantly above the national average.

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

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

Not Applicable

**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*)

Not Applicable

**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*)

Not Applicable

**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*)

Not Applicable

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

Not applicable. The measure has been collected since 2008 and hospitals transmitting data with missing data on any of the critical data elements are not accepted.

**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*)

Not applicable.

**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*)

Not applicable.

**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*)

Not applicable.