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

**Measure Number** (*if previously endorsed*)**:** N/A

**Measure Title**: Hospital Harm - Opioid-Related Adverse Events

**Date of Submission**: TBD

**Type of Measure:**

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

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

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

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

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

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

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

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

We partnered with a quality measure reporting service provider and an academic health center in Northern California to support alpha testing of the measure, in particular, the high-level measure feasibility assessment. We then continued the partnership with the quality measure reporting service provider and engaged a different health system to complete the implementation testing of the final measure specification exported from the measure authoring tool (MAT). Measure implementation testing occurred in six test sites across two different EHR vendors/systems—Cerner and Meditech.

Alpha testing aimed to determine if test sites can capture the critical data elements used in the measure based on codes in the updated value sets. Beta testing, on the other hand, is larger in scope, and consisted of two phases, assessing the measure feasibility in detail and assessing the measure’s scientific acceptability.

In phase 1 of beta testing, we conducted a detailed feasibility assessment of measure implementation. Specifically, we surveyed participant test sites on the extent to which critical data elements required for measure implementation are available in their EHR systems in a structured format, from an authoritative source, coded using nationally recognized terminologies or could be mapped, and the extent of impact on current clinical and technical workflows. During this phase of testing, we assessed measure feasibility in detail using 23 participant test sites across four EHR vendors/systems—Epic, Cerner, Allscripts, and Meditech.

In phase 2 of beta testing, we assessed the measure’s scientific acceptability using the patient EHR data extracted from six implementation test sites and calendar year 2019. Cerner and Meditech systems were equally represented across the 6 implementation test sites. We also drew a sub-sample of patient data for each of the six implementation test sites and performed a parallel-form comparison, comparing data extracted from the EHR electronically to data manually abstracted from patient medical records.

The dataset used varies by testing type. Please see section 1.7 for details.

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

Alpha testing used data from calendar year 2018 and phase 2 of beta testing used data from calendar year 2019. Please see section 1.7 for details.

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

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

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

Five hospitals participated in alpha testing (in this measure, data queries). A total of 23 hospitals participated in phase 1 of beta testing, or detailed feasibility assessment, and six hospitals participated in phase 2 of beta testing, or implementation testing. Participant test sites vary by EHR vendor systems, bed size, geographic location, teaching/non-teaching status, and urban/rural representation. Please see section 1.7 for details.

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

Alpha testing focused on measure feasibility at a high-level, with the results ultimately informing the measure specification development. Hence, we did not use patient or encounter level data in alpha testing. We instead queried five hospitals’ EHR databases to determine whether or not they can capture the critical data elements required for measure implementation based on codes in the updated measure value set.

Phase 1 of beta testing assessed in detail the feasibility of measure implementation, and the goal was to determine, within each test site’s EHR system, if the critical data elements are: 1) readily available in a structured format; 2) from an authoritative source and/or highly likely to be correct; 3) coded in a nationally accepted terminology standard or can be mapped to that terminology standard; and 4) routinely collected as part of clinical care and require no or limited additional data entry from a clinician or other providers, and no EHR interface changes are needed. To that end, we designed a web-based questionnaire via the SurveyMonkey® platform and distributed the questionnaire to test participants. No patient or encounter level data were used in phase 1 of beta testing, which included 23 test sites.

Phase 2 of beta testing assessed the measure’s scientific acceptability and employed one year (1/1/2019 – 12/31/2019) of patient EHR data from six implementation test sites. A total of 1,537, 1,889, 1,544, 9,413, 10,827, and 15,261 unique patients (inclusive of ED visits or observation stay that eventually admitted to hospitals for inpatient treatment, or inpatient only) were extracted from the six implementation test sites, respectively. These patients corresponded to 1,839, 2,089, 1,784, 11,273, 13,307, and 18,425 measure denominator encounters or qualified admissions. The average age of patients in the measure denominator ranged from 51 to 61, and over half of them were female and White. No diagnosis information was extracted as measure implementation does not require such information.

Parallel-form comparison (comparing electronically extracted EHR data to manually abstracted data from the same patient’s medical record) was based on a randomly selected sub-sample from the measure initial population in each of the six implementation test sites. Specifically, for each of the six sites, we randomly sampled 100 patient encounters from the measure initial population, while holding fixed the distribution of patient demographic characteristics (age, sex, race/ethnicity, and primary payer) in the full sample. We used random sampling without replacement.

Please see section 1.7 for details on the dataset used for the different aspects of measure testing.

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

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

**Table 1. Dataset Descriptions**

| **Dataset** | **Applicable Section in the Testing Attachment** | **Description of Dataset** | **EHR Vendor** |
| --- | --- | --- | --- |
| **Alpha** | N/A. The high-level analysis of measure feasibility used drug frequency counts based on the RxNorm codes in the updated measure value set. | Dates of Data: 1/1/2018 to 12/31/2018  Number of hospitals: 5  Four hospitals associated with the quality measure reporting service provider use Meditech, and one academic medical center located in Northern California use Epic. | Meditech and Epic |
| **Beta Dataset – Feasibility Assessment** | N/A. We surveyed 23 hospitals about the feasibility of implementing the measure as specified within their EHR systems. The survey data laid the foundation for the measure’s feasibility scorecard | Dates of Data: survey took place between February and May 2020  Number of Hospitals: 23  EHR vendors/systems (number of test sites):   * Epic (5) * Cerner (5) * Allscripts (3) * Meditech (10)   Bed size (number of test sites)::   * 25-99 beds (5) * 100-199 beds (8) * 200-499 beds (6) * 500+ beds (4)   Census region (number of test sites)::   * Northeast (2) * Midwest (7) * South (6), * West (8)   Urban/rural Area (number of test sites)::   * Urban (12) * Rural (11) | Epic, Cerner, Meditech, and Allscripts |
| **Beta Dataset – Implementation Testing** | Section 2a2 Reliability Testing  Section 2b1 Validity Testing  Section 2b4 Identification of Statistically Significant and Meaningful Differences in Performance  Section 2b6 Missing Data Analysis | Dates of Data: 1/1/2019 – 12/31/2019  Number of Hospitals: 6  Number of Admissions (or measure denominator encounters):   * Hospital 1: 1,839 * Hospital 2: 2,089 * Hospital 3: 1,784 * Hospital 4: 11,273 * Hospital 5: 13,307 * Hospital 6: 18,425   Number of Unique Patients (or unique patients from the measure denominator encounters):   * Hospital 1: 1,537 * Hospital 2: 1,889 * Hospital 3: 1,544 * Hospital 4: 9,413 * Hospital 5: 10,827 * Hospital 6: 15,261   For Validity Testing: randomly selected sample of 100 qualified admissions for each of the six implementation test sites.  The six implementation test sites vary by bed size (between 71 to over 500 beds), teaching and non-teaching status, are in five different states, and half of them are located in urban areas. Test sites 1 to 3 use Meditech and test sites 4 to 6 use Cerner. | Meditech and Cerner |

**Patient descriptive characteristics in Beta Dataset – Implementation Testing are as follows:**

* Average age at admission across the six implementation test sites ranged between 51 and 61
* Gender distribution at admission across the six implementation test sites ranged from 58% female to 68% female or 32% male to 42% male
* Race distribution at admission across the six implementation test sites ranged from 62% white to 96% white

Detailed patient descriptive characteristics included in the analysis by implementation test site for **the second phase of beta testing, or implementation testing** are provided in **Tables 2 and 3**.

Table 2. Demographic Characteristics of Measure Initial Population (Site 1-3)

| **Initial Population Characteristics** | **Test Site 1** | | **Test Site 2** | | **Test Site 3** | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **n** | **%** | **n** | **%** | **n** | **%** |
| Number of patient-encounters | 1,839 | 100% | 2,089 | 100% | 1,784 | 100% |
| Number of unique patients | 1,537 | 100% | 1,889 | 100% | 1,544 | 100% |
| Age Mean (Std.Dev) | 55.1 (22.0) |  | 58.0 (21.6) |  | 60.7 (20.2) |  |
| Age bins |  |  |  |  |  |  |
| 18-35 | 429 | 28% | 415 | 22% | 271 | 18% |
| 36-64 | 500 | 33% | 629 | 33% | 473 | 31% |
| 65+ | 607 | 39% | 845 | 45% | 799 | 52% |
| Sex |  |  |  |  |  |  |
| Male | 513 | 33% | 612 | 32% | 568 | 37% |
| Female | 1,024 | 67% | 1,277 | 68% | 976 | 63% |
| Race |  |  |  |  |  |  |
| Black or African American | 39 | 3% | 80 | 4% | 15 | 1% |
| White | 1,483 | 96% | 1,795 | 95% | 1,487 | 96% |
| Other | 15 | 1% | 13 | 1% | 40 | 3% |
| Unknown | 0 | 0% | 1 | 0% | 2 | 0% |
| Ethnicity |  |  |  |  |  |  |
| Hispanic or Latino | 0 | 0% | 74 | 4% | 10 | 1% |
| Non-Hispanic | 0 | 0% | 1,810 | 96% | 1,534 | 99% |
| Unknown | 1,537 | 100% | 5 | 0.3% | 0 | 0% |
| (Primary) Payer |  |  |  |  |  |  |
| Medicare | 718 | 47% | 1,025 | 54% | 849 | 55% |
| Medicaid | 395 | 26% | 377 | 20% | 191 | 12% |
| Private Insurance | 292 | 19% | 314 | 17% | 456 | 30% |
| Self-pay or Uninsured | 1 | 0% | 19 | 1% | 12 | 1% |
| Other† | 116 | 8% | 62 | 3% | 20 | 1% |
| Unknown | 15 | 1% | 92 | 5% | 16 | 1% |

Note: n = frequency, % = percentage, and Std.Dev = standard deviation. † “Other” include other government plans (e.g. federal, state, local) than Medicare and Medicaid, Worker’s Compensation plans, or other unspecified plans.

Table 3. Demographic Characteristics of Measure Initial Population (Site 4-6)

| **Initial Population Characteristics** | **Test Site 4** | | **Test Site 5** | | **Test Site 6** | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **n** | **%** | **n** | **%** | **n** | **%** |
| Number of patient-encounters | 11,273 | 100% | 13,307 | 100% | 18,425 | 100% |
| Number of unique patients | 9,413 | 100% | 10,827 | 100% | 15,261 | 100% |
| Age Mean (Std.Dev) | 50.5 (20.7) |  | 56.1 (19.8) |  | 51.3 (19.1) |  |
| Age bins |  |  |  |  |  |  |
| 18-35 | 3,123 | 33% | 2,361 | 22% | 4,267 | 28% |
| 36-64 | 3,418 | 36% | 4,415 | 41% | 6,316 | 41% |
| 65+ | 2,861 | 30% | 4,037 | 37% | 4,653 | 30% |
| Sex |  |  |  |  |  |  |
| Male | 3,226 | 34% | 4,412 | 41% | 6,473 | 42% |
| Female | 6,187 | 66% | 6,415 | 59% | 8,788 | 58% |
| Race |  |  |  |  |  |  |
| Black or African American | 564 | 6% | 780 | 7% | 1,368 | 9% |
| White | 7,034 | 75% | 6,708 | 62% | 12,002 | 79% |
| Other | 1,514 | 16% | 1,459 | 13% | 1,304 | 9% |
| Unknown | 301 | 3% | 1,880 | 17% | 587 | 4% |
| Ethnicity |  |  |  |  |  |  |
| Hispanic or Latino | 4,033 | 43% | 984 | 9% | 4,182 | 27% |
| Non-Hispanic | 5,140 | 55% | 7,832 | 72% | 9,853 | 65% |
| Unknown | 240 | 3% | 2,011 | 18.6% | 1,226 | 8% |
| (Primary) Payer |  |  |  |  |  |  |
| Medicare | 2,174 | 23% | 2,900 | 27% | 5,399 | 35% |
| Medicaid | 4,729 | 50% | 4,471 | 41% | 5,121 | 34% |
| Private Insurance | 1,206 | 13% | 1,859 | 17% | 3,600 | 24% |
| Self-pay or Uninsured | 1,093 | 12% | 1,348 | 12% | 524 | 3% |
| Other | 210 | 2% | 249 | 2% | 617 | 4% |
| Unknown | 1 | 0% | 0 | 0% | 0 | 0% |

Note: n = frequency, % = percentage, and Std.Dev = standard deviation. † “Other” include other government plans (e.g. federal, state, local) than Medicare and Medicaid, Worker’s Compensation plans, or other unspecified plans.

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

As described in section 1.7 and **Tables 2 and 3**, we collected information on the following social risk factors using data extracted from hospital EHR systems: race, ethnicity, and primary payer.

**2a2. RELIABILITY TESTING**

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

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure** (e.g., *signal-to-noise analysis*)

**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

We assessed data element reliability using two methods. First, we calculated the rate of missing or erroneous data for all critical data elements required for measure implementation. Specifically, for each of the six implementation test sites, we tabulated the number of measure denominator encounters where critical data elements are either missing or showing erroneous values. Second, we used Cohen’s Kappa to quantify the inter-rater agreement. For the second approach, we turned to six sub-samples, each of which consists of 100 patient encounters drawn from the corresponding site’s measure initial population via random sampling without replacement. The number of observation (100) is based on a formula that we describe in detail in section 2b1.

Cohen’s Kappa can be conceptualized in a stylized matrix, as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Rater A | Rater B | | Total |
| 1 | 2 |
| 1 |  |  |  |
| 2 |  |  |  |
| Total |  |  |  |

, where Rater A can be viewed as the implementation test site’s certified electronic health record technology (CEHRT) test environment and Rater B as the clinical abstractor (e.g., clinician informaticist). If we define the proportion of agreement expected by chance as:

, then Cohen’s Kappa coefficient is equal to , where denotes the observed proportion of agreement between the two raters.

Measure Score Reliability

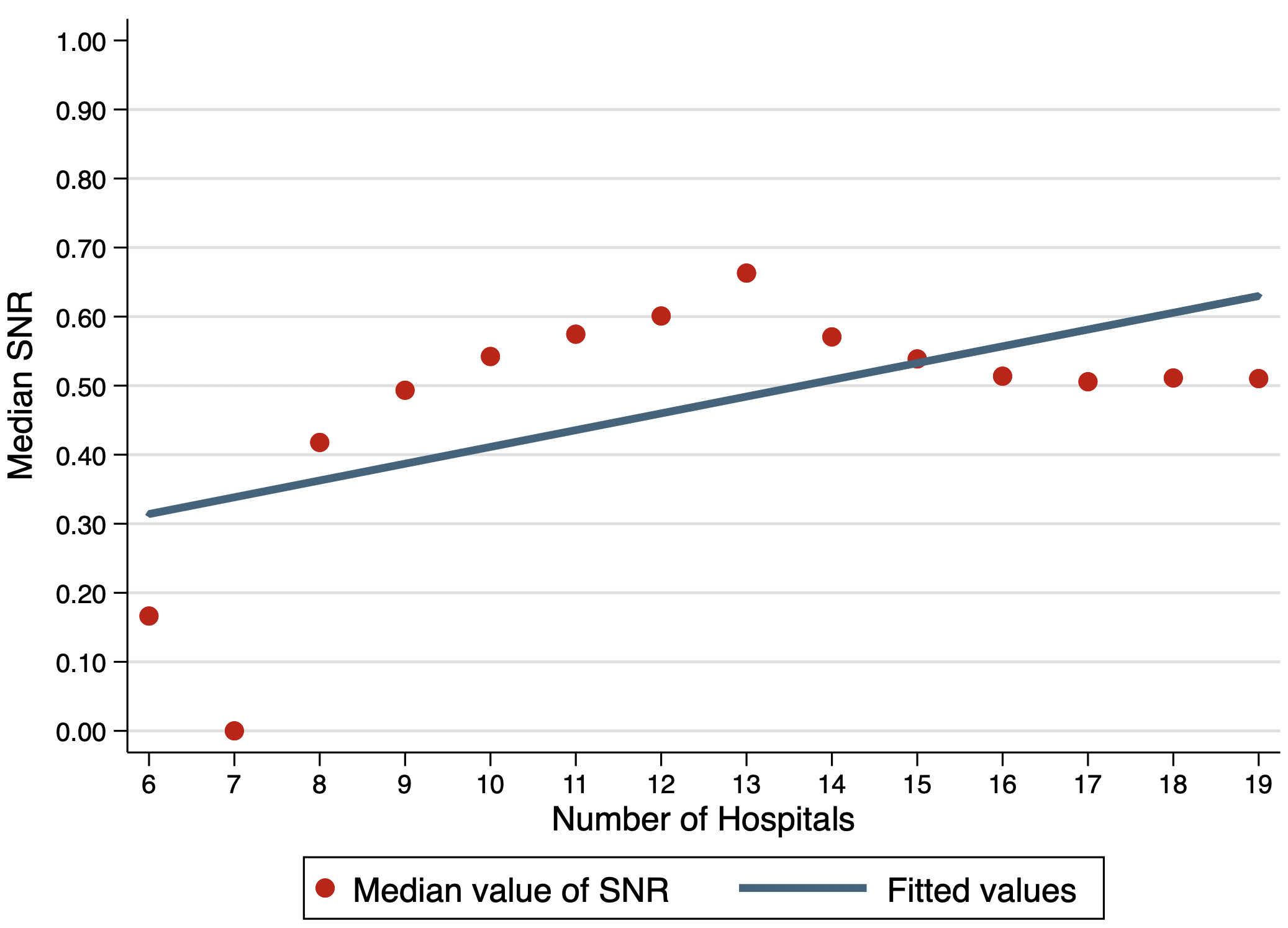
Measure score reliability describes how well one can confidently distinguish the performance of one measured entity (e.g., hospital) from another, and is typically evaluated based on signal-to-noise ratios (SNR) calculated via John Adams’ beta-binomial method (Adams, 2009). This method, however, requires a sufficient number of measured entities (hospitals in this case) in order to obtain consistently estimated alpha and beta parameters, which are then to form the hospital-level variance (or signal). Because only six hospitals (two different EHR systems) participated in implementation testing, it is conceivable that the alpha and beta parameters will not be estimated with precision and the end result could be spurious. Compounding the data limitation issue is the large coefficient of variation for the measure performance rates in the six implementation test sites. This suggests that one would need an even larger amount of data (number of hospitals) to clearly distinguish the hospital-level variance (signal) from the within-hospital variance (noises). For these reasons, we did not perform measure score reliability assessment.

However, to evaluate if a larger set of data (number of hospitals) can yield a higher SNR, we worked with the health system that participated in measure implementation testing and collected high-level numerator and denominator counts based on the measure specification from a total of 16 hospitals. We then randomly selected a subset of hospitals from this pool, starting from three hospitals to the full set of 16 hospitals, and calculated SNR for each hospital using the Adams’ beta-binomial method. Our baseline dataset is thus comprised of six hospitals, which are the three hospitals associated with the quality reporting service provider and the three randomly selected ones.

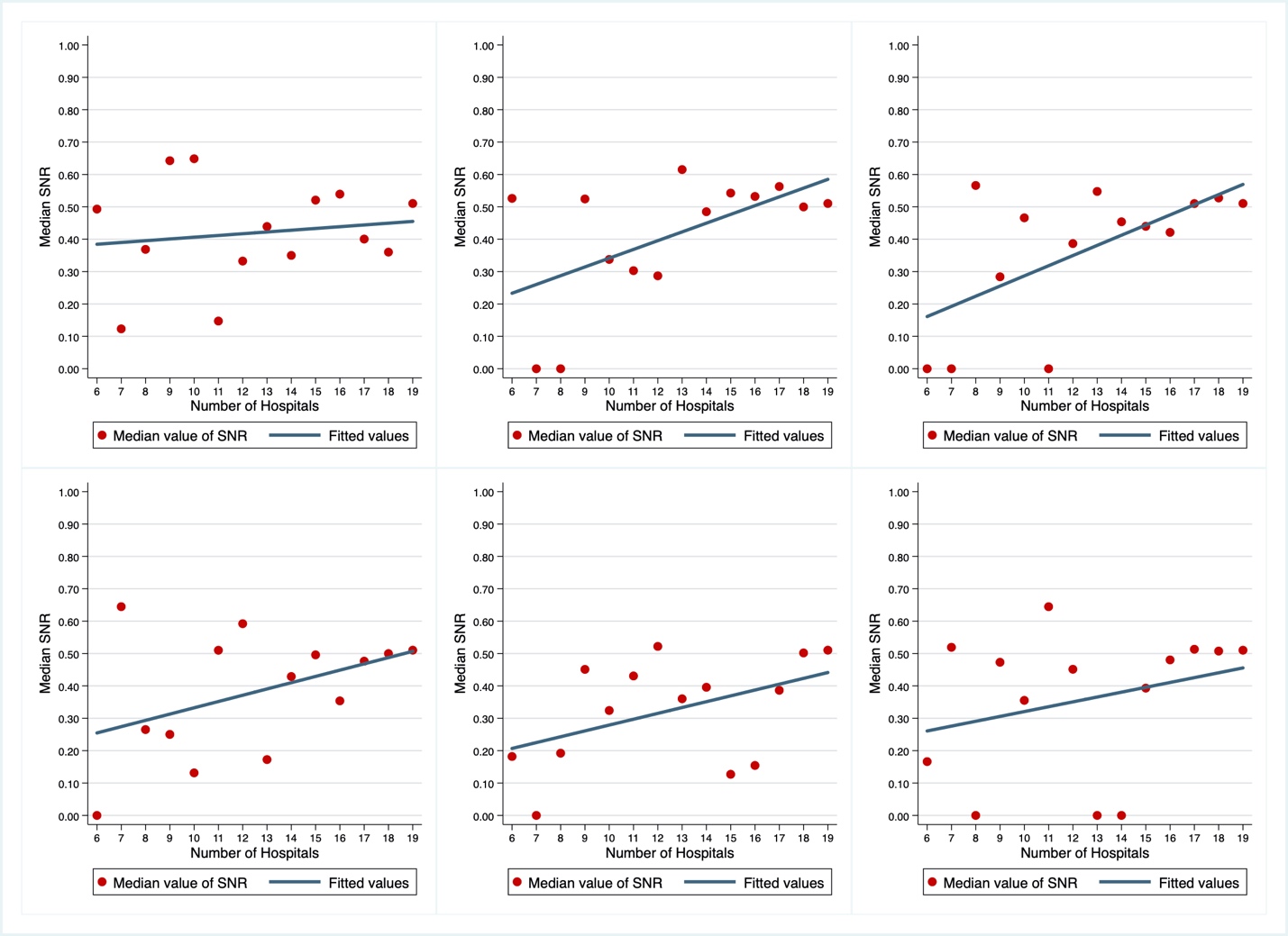
Figure 1 (attached in the intent-to-submit form) presents a scatterplot of the median value of SNR estimated across the hospitals and a best-fit line depicting the general tendency of SNR as more hospitals are used for estimation. The horizontal axis denotes the number of hospitals that contributed data. Three points are worth noting. First, the scatterplot is only one of all the permutated scenarios; therefore, readers should not take the value (position) of each dot as final. Second, the scatterplot varies from one estimation to another, but the best-fit line is always upward trending, suggesting that there exists a positive relationship between the number of hospitals used for estimation and the value of SNR (Figure 2, which is also attached in the intent-to-submit form). Third, the variability of SNRs (dots) provides supporting evidence to our rationale for not assessing the measure score level reliability. Evidently, the conclusion one may draw from the SNR analysis can be far from definitive when only a few data points are used for estimation.

Of note, the test sample used in the measure implementation testing has satisfied the NQF Measure Evaluation Criteria and Guidance.

**Figure 1. Relationship Between the Number of Hospitals and the Median Value of SNR**



**Figure 2. Relationship Between the Number of Hospitals and the Median Value of SNR**



References:

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

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Data Element Reliability

**Tables 4 and 5** display the data element level reliability, evaluated on the basis of percentage of missing or erroneous data for every critical data element needed for measure implementation. The results suggest that all critical data elements are reliably and consistently captured in patient EHRs.

**Table 4. Data Element Reliability Results (Frequency of Missing or Erroneous Data) for the Critical Data Elements (Sites 1-3)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Test Site 1** | | | **Test Site 2** | | | **Test Site 3** | | |
| Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) |
| Patient inpatient encounter discharge DateTime | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Patient birth date | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Opioid administration DateTime | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Naloxone Administration DateTime | 0 | 2 | 0% | 0 | 7 | 0% | 0 | 8 | 0% |
| Naloxone Administration Location | 0 | 2 | 0% | 0 | 7 | 0% | 0 | 8 | 0% |

**Table 5. Data Element Reliability Results (Frequency of Missing or Erroneous Data) for the Critical Data Elements (Sites 4-6)**

| **Data Element** | **Test Site 4** | | | **Test Site 5** | | | **Test Site 6** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) |
| Patient inpatient encounter discharge DateTime | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Patient birth date | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Opioid administration DateTime | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Naloxone Administration DateTime | 0 | 51 | 0% | 0 | 44 | 0% | 0 | 64 | 0% |
| Naloxone Administration Location | 0 | 51 | 0% | 0 | 44 | 0% | 0 | 64 | 0% |

**Table 6** shows the Kappa coefficients calculated for each critical data element and for each of the six implementation test sites. Except for a discrepancy clinical abstractors found in test site 5 that yielded the Kappa coefficient equal to 0.98, all the other Kappa coefficients are 1, indicating perfect agreement. As noted in section 2a2.2, we calculated Kappa coefficients based on six randomly selected sub-samples and by comparing electronically extracted EHR data to manually abstracted data from the same patient’s medical record (parallel-form comparison). The misaligned case identified during this process of parallel-form comparison pointed to a numerator event even though the quality reporting engine marked it as denominator only. In section 2b1, we provide details to this misaligned case and discuss lessons learned from the parallel-form comparison.

**Table 6. Data Element Reliability Results (Cohen’s Kappa) for the Critical Data Elements**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Test Site 1** | **Test Site 2** | **Test Site 3** | **Test Site 4** | **Test Site 5** | **Test Site 6** |
| Kappa | Kappa | Kappa | Kappa | Kappa | Kappa |
| Patient have an inpatient encounter with a discharge date between 1/1/19 and 12/31/19 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Patient age ≥ 18 at the start of the encounter | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| An opioid was administered to the patient during the encounter | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| An opioid antagonist was administered to the patient during the encounter | 1.0 | 1.0 | 1.0 | 1.0 | 0.98 | 1.0 |
| An opioid antagonist was administered to the patient both within 12hrs of the opioid administration and outside of the operating room | 1.0 | 1.0 | 1.0 | 1.0 | 0.98 | 1.0 |

**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?*)

**Tables 4 through 6** suggest that all critical data elements are reliably and consistently captured in patient EHRs and that there is a strong concordance between data extracted from the EHR electronically and data extracted from patient medical records manually (“gold standard”).

Our interpretation of the Kappa coefficient is based on standards established by Viera and Garrett (2005):

• 0.4 – 0.6: moderate agreement

• 0.6 – 0.8: substantial agreement

• 0.8 – 1: almost perfect agreement

Reference:

Viera, Anthony J., and Joanne M. Garrett. "Understanding interobserver agreement: the kappa statistic." *Fam Med* 37, no. 5 (2005): 360-363.

**2b1. VALIDITY TESTING**

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

**Performance measure score**

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

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

Data Element Validity

We assessed data element level validity by evaluating the agreement rate of electronically extracted data elements from patient EHR and manually chart abstracted data elements from the same patient’s medical record. To define agreement, we considered each data element matched if the electronically extracted value exactly matched the manual abstraction value. For example, if a patient deemed 18 or above at the start of the encounter according to information in the EHR was indeed 18 or above based on the clinical review of his/her medical record (date of birth and the encounter start date), then we treated this case as fully matched. Or, if the EHR indicated that the patient received an opioid antagonist (naloxone) within 12 hours of the prior opioid and outside of the operating room suite, and the clinical abstractor confirmed both the timing and location of administration in the patient’s medical record, then we treated this case as fully matched.

Measure Score Validity

To assess measure score level validity, we turned to four statistics: positive predictive value, sensitivity, negative predictive value, and specificity. Positive predictive value, or PPV, describes the probability that a patient with a positive result reported by the EHR is also a positive result confirmed by the clinical abstraction. In the context of the current measure, PPV is the probability that a EHR-reported ORAE is a valid ORAE based on the clinical review of the patient’s medical record. Sensitivity describes the probability that a patient with an ORAE based on the medical abstraction is correctly classified as an ORAE by the EHR. Negative predictive value, or NPV, describes the probability that a patient with a negative result in the EHR is also a negative result based on the clinical abstraction. In the current measure, NPV thus denotes the probability that an at-risk patient who did not have an EHR-reported ORAE was also not an ORAE based on the clinical abstraction. Specificity, mirroring the relationship between PPV and sensitivity, describes the probability that a patient who is not an ORAE based on the clinical abstraction is correctly classified as a non-ORAE by the EHR.

We acknowledge that PPV, sensitivity, NPV, and specificity are by no means the only metrics measure developers utilize to assess the measure score validity. For example, measure convergent validity is another commonly used method in claims-based measures. However, we did not adopt this approach because the number of hospitals participated in measure implementation testing is too few to render this approach meaningful. As a reminder, only six hospitals participated in measure implementation testing. In addition, for measures that count harm events without other statistical manipulation, such as regression-based risk adjustments, the confirmation that the measure logic is accurately capturing the true harm event is the gold standard for assessing the measure score validity.

Differing from the measure score reliability assessment where we used the full sample, measure validity testing was based upon a random sample of 100 patient encounters (both numerator and denominator-only cases). We calculated this minimum required sample size (MRSS, 100 encounters) using PPV as the primary endpoint, and approximated MRSS using the one-sample proportion formula:

, where denotes the type I error rate, denotes the margin of error, and is PPV. We simulated a series of and target values for MRSS and 95% confidence interval (CI). For example, with a of 6% and a target PPV of 0.9, MRSS equal to 100 produced a 95% CI of PPV equal to 0.84 – 0.96. We thus believe that MRSS equal to 100 can produce an accurate PPV estimation.

For each of the six implementation test sites, we randomly drew 100 patient encounters from the measure initial population, meanwhile holding fixed the distribution of patient demographic characteristics (age, sex, race/ethnicity, and primary payer) in the full sample. We then manually reviewed patient medical records for each of the sampled patients, compared abstraction data to what were extracted from their EHRs, and used these data to calculate the PPV, sensitivity, NPV, and specificity. We used random sampling without replacement.

In our full sample, the total number of numerator cases in implementation test sites 1 through 5 are 2, 7, 8, 51, and 44, respectively. We thus included all these numerator cases in their corresponding 100 patient encounters or sub-samples. Test site 6 has a total of 64 numerator cases, and we randomly selected 50 cases in order to maintain balance between the numerator and denominator-only cases. **Table 7** shows the number of numerator and denominator-only cases included in each of the six sub-samples for the parallel-form comparison.

**Table 7. Number of Numerator and Denominator-only Cases Included in the Randomly Selected 100 Patient Encounters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Test Site 1** | **Test Site 2** | **Test Site 3** | **Test Site 4** | **Test Site 5** | **Test Site 6** |
| Count | Count | Count | Count | Count | Count |
| Numerator | 2 | 7 | 8 | 51 | 44 | 50 |
| Denominator-only | 98 | 93 | 92 | 49 | 56 | 50 |

Manual abstraction was performed by the experienced medical record reviewers. We provided them with a guidance document and an Excel workbook to document findings in the sampled patient’s medical record. We pre-populated Excel workbooks with the unique patient identifiers only and instructed abstractors to input all the other data, including free text and summary notes, from the patient EHRs and medical records.

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

Data Element Validity

**Tables 8 through 10** display the agreement rate between data electronically extracted from the sampled patients’ EHRs and data manually abstracted from their medical records, for all critical data elements used in the measure.

**Table 8. Data Element Validity Results (Agreement Rate) for the Critical Data Elements (Sites 1 and 2)**

| **Data Element** | **Test Site 1** | | | **Test Site 2** | | |
| --- | --- | --- | --- | --- | --- | --- |
| # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) | # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) |
| Patient have an inpatient encounter with a discharge date between 1/1/19 and 12/31/19 | 100 | 100 | 100% | 100 | 100 | 100% |
| Patient age ≥ 18 at the start of the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid was administered to the patient during the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid antagonist (naloxone) was administered to the patient during the encounter | 2 | 2 | 100% | 7 | 7 | 100% |
| An opioid antagonist (naloxone) was administered to the patient both within 12hrs of the opioid administration and outside of the operating room | 2 | 2 | 100% | 7 | 7 | 100% |

**Table 9**. **Data Element Validity Results (Agreement Rate) for the Critical Data Elements (Sites 3 and 4)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Test Site 3** | | | **Test Site 4** | | |
| # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) | # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) |
| Patient have an inpatient encounter with a discharge date between 1/1/19 and 12/31/19 | 100 | 100 | 100% | 100 | 100 | 100% |
| Patient age ≥ 18 at the start of the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid was administered to the patient during the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid antagonist (naloxone) was administered to the patient during the encounter | 8 | 8 | 100% | 51 | 51 | 100% |
| An opioid antagonist (naloxone) was administered to the patient both within 12hrs of the opioid administration and outside of the operating room | 8 | 8 | 100% | 51 | 51 | 100% |
|  |  |  |  |  |  |  |

**Table 10**. **Data Element Validity Results (Agreement Rate) for the Critical Data Elements (Sites 5 and 6)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Test Site 5** | | | **Test Site 6** | | |
| # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) | # Cases Matched in EHR (n) | # Cases in Abstraction (n) | Percent Match (%) |
| Patient have an inpatient encounter with a discharge date between 1/1/19 and 12/31/19 | 100 | 100 | 100% | 100 | 100 | 100% |
| Patient age ≥ 18 at the start of the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid was administered to the patient during the encounter | 100 | 100 | 100% | 100 | 100 | 100% |
| An opioid antagonist (naloxone) was administered to the patient during the encounter | 44 | 45 | 98% | 50 | 50 | 100% |
| An opioid antagonist (naloxone) was administered to the patient both within 12hrs of the opioid administration and outside of the operating room | 44 | 45 | 98% | 50 | 50 | 100% |

Across the six implementation test sites, all but two data elements showed an agreement rate of 100%, indicating that valid and accurate data were extracted from patients’ EHRs. The misaligned case in test site 5 was due to the clinical abstractor review determining the encounter to be a numerator event, yet EHR data suggested it to be in the measure denominator only. In this particular situation, the facility uses a paper-based anesthesia record when documenting operation room (OR)-specific medication administration. During the encounter in question, the patient received an opioid inside of the OR and later an opioid outside of the OR suite. Naloxone was administered within 12 hours of the OR administered opioid (numerator qualifying event) but prior to the second opioid administration. Because the opioid inside of the OR was not electronically retrievable, the quality reporting engine was only able to capture the encounter in the measure denominator. That is, an opioid was received during the encounter but not within the 12 hours prior to the naloxone administration.

One may have concern that this clinical workflow may cause the measure to suffer from false negatives. While we cannot eliminate that the measure is immune to false negatives, we found supportive evidence indicating that the concern is not grave. In particular, during the manual abstraction process, we learned that test sites 4, 5, and 6 (all part of a single hospital system) share the same medication documentation pattern inside of the OR, and that is all OR-specific medication administrations are documented in paper-based anesthesia records. Across the 155 (49 + 56 + 50) denominator-only cases from test sites 4 to 6, we only saw one false negative. The low rate of false negative (0.6%) provides some degree of confidence that the issue is not widely seen in the harm event the current measure seeks to identify. Moreover, for hospitals (such as test sites 1 to 3) that utilize electronic medication administration records (eMARs) throughout, such false negative is eliminated.

To examine if the numerator cases identified by the quality reporting engine are true positives, clinical abstractors pulled additional information regarding the indication for and subsequent reaction to the naloxone administration from the nurse notes and physician orders. We grouped patient responses to naloxone administration as follows: 1) patient showed clear signs of reaction after the naloxone administration; 2) patient showed little signs of reaction; and 3) patient responses were not documented. The first group encompasses scenarios ranging from “patients became less drowsy” to “patients woke up immediately after naloxone administration.” Group 2 includes scenarios such as “patient mentation changed slightly after naloxone administration” and “patients had little improvement after naloxone administration.” Figure 3 (attached in the intent-to-submit form) plots the response distribution by test site and Figure 4 (attached in the intent-to-submit form) shows the response distribution after we pool the data from all test sites.

**Figure 3. Patient Responses to Naloxone Administration by Test Site**

Figure 3 plots the distribution of patient's responses to naloxone administration within each of the six beta implementation test sites.  

Notes: value 1 indicates patients showed clear signs of reactions after naloxone administration, value 2 indicates patients showed little signs of reactions, and value 3 indicates that patient responses were not documented.

**Figure 4. Patient Responses to Naloxone Administration Across Six Test Sites**

Figure 4 shows the distribution of patient responses to naloxone administration after we pool the data from all test sites.

Notes: value 1 indicates patients showed clear signs of reactions after naloxone administration, value 2 indicates patients showed little signs of reactions, and value 3 indicates that patient responses were not documented.

By excluding “no responses were documented” from the group, 76% of the reviewed numerator cases had nurse notes indicating that patients showed clear signs of reaction after the naloxone. The most frequently documented response was that “patients became more awake.” This qualitative piece of evidence solidifies our evaluation of measure logic and suggests that the measure can correctly predict a true positive (excessive opioid administration or ORAE).

The remaining 24%, where patients showed little signs of reaction after the naloxone administration, may still cause concerns for false positives. We caution that patients showing no immediate responses may be due to the inadequate dosage of naloxone, as there were a few instances identified during the manual abstraction where patients became responsive only after the second naloxone.

We also found that some test sites have used, though not consistently, the Pasero Opioid-induced Sedation Scale (POSS) in recording the appropriateness of opioid dosage. POSS typically consists of 5 scales:

| POSS | Interpretation |
| --- | --- |
| S = Sleep, easy to arouse | Acceptable; no action necessary; may increase opioid dose if needed |
| 1 = Awake and alert | Acceptable; no action necessary; may increase opioid dose if needed |
| 2 = Slightly drowsy, easily aroused | Acceptable; no action necessary; may increase opioid dose if needed |
| 3 = Frequently drowsy, arousable, drifts off to sleep during conversation | Unacceptable; monitor respiratory status and sedation level closely until sedation level is stable at less than 3 and respiratory status is satisfactory; decrease opioid dose 25% to 50% or notify prescriber or anesthesiologist for orders; consider administering a non-sedating, opioid-sparing nonopioid, such as acetaminophen or a NSAID, if not contraindicated. |
| 4 = Somnolent, minimal or no response to verbal and physical stimulation | Unacceptable; stop opioid; consider administering naloxone; notify prescriber2 or anesthesiologist; monitor respiratory status and sedation level closely until sedation level is stable at less than 3 and respiratory status is satisfactory. |

Of the identified numerator cases where POSS were used, most showed an initial POSS of 3 or 4. After the naloxone administration, patients’ POSS decreased to 1 or 2. We believe that leveraging POSS in classifying measure numerator event can further increase the accuracy in predicting true positives (ORAEs). But we underscore that the use of POSS is not universal across the six test sites. Moreover, among those who use POSS, the utilization is inconsistent.

Measure Score Validity

**Tables 11-14** show the measure score level validity, evaluated by PPV, sensitivity, NPV, and specificity. As mentioned in section 2b1.2, we define PPV as the probability that an EHR-reported ORAE is a valid ORAE based on the clinical review of patients’ medical records. We define sensitivity as the probability that a patient had an ORAE based on the medical record was correctly classified by the EHR as having an ORAE. We define NPV and specificity accordingly. Each component of the measure was validated by the clinical abstractors, and we evaluated the overall agreement between data in the EHR and data in the medical record.

Denominator PPV, assessing the percent of patient encounters that correctly belong to the measure denominator, is 100% for all test sites except test site 5, where the denominator PPV is 98%. Numerator PPV is 100% for all six test sites. Sensitivity is 100% in all but one test site and specificity is 100% in all sites.

**Table 11. Measure Score Validity (PPV) for the Sampled Patient Encounters (Sites 1-3)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure Component** | **Test Site 1** | | | **Test Site 2** | | | **Test Site 3** | | |
| Positive in Chart Abstraction | Positive in EHR Data | PPV | Positive in Chart Abstraction | Positive in EHR Data | PPV | Positive in Chart Abstraction | Positive in EHR Data | PPV |
| Initial population | 100 | 100 | 100% | 100 | 100 | 100% | 100 | 100 | 100% |
| Denominator only | 98 | 98 | 100% | 93 | 93 | 100% | 92 | 92 | 100% |
| Numerator | 2 | 2 | 100% | 7 | 7 | 100% | 8 | 8 | 100% |

**Table 12. Measure Score Validity (PPV) For the Sampled Patient Encounters (Sites 4-6)**

| **Measure Component** | **Test Site 4** | | | **Test Site 5** | | | **Test Site 6** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Positive in Chart Abstraction | Positive in EHR Data | PPV | Positive in Chart Abstraction | Positive in EHR Data | PPV | Positive in Chart Abstraction | Positive in EHR Data | PPV |
| Initial population | 100 | 100 | 100% | 100 | 100 | 100% | 100 | 100 | 100% |
| Denominator only | 49 | 49 | 100% | 55 | 56 | 98% | 50 | 50 | 100% |
| Numerator | 51 | 51 | 100% | 45 | 44 | 100% | 50 | 50 | 100% |

**Table 13. Measure Score Validity (Sensitivity, NPV, and Specificity) for the Sampled Patient Encounters (Sites 1-3)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Test Site 1 (N=100)** | | | **Test Site 2 (N=100)** | | | **Test Site 3 (N=100)** | | |
| Sensitivity | NPV | Specificity | Sensitivity | NPV | Specificity | Sensitivity | NPV | Specificity |
| **ORAE** | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

**Table 14. Measure Score Validity (Sensitivity, NPV, and Specificity) for the Sampled Patient Encounters (Sites 4-6)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Test Site 4 (N=100)** | | | **Test Site 5 (N=100)** | | | **Test Site 6 (N=100)** | | |
| Sensitivity | NPV | Specificity | Sensitivity | NPV | Specificity | Sensitivity | NPV | Specificity |
| **ORAE** | 100% | 100% | 100% | 98% | 98% | 100% | 100% | 100% | 100% |

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Data Element Validity

Across the six implementation test sites, all but two data elements showed a match rate of 100%, indicating that valid and accurate data were extracted from patient EHRs. The exceptions in test site 5 were due to a documentation preference. As we discussed in section 2b1.3, across the 155 (49 + 56 + 50) denominator-only cases from test sites 4 to 6 who share the same documentation pattern inside of the OR, we found only one misaligned case. The low false negative rate provides some degree of confidence that the issue is not widely seen in the harm event the current measure seeks to identify. Moreover, for hospitals that utilize eMARs throughout, this misalignment will be eliminated. Because all hospital-based EHR vendor systems offer anesthesia modules that can document medication electronically, there should be no technical limitation in transitioning from paper-based documentation to electronic documentation.

Measure Score Validity

Across the six implementation test sites PPV is 100%, suggesting that in all cases the qualified admissions have met the criteria for a ORAE in both the chart-abstracted and EHR-extracted data.

Sensitivity is 100% in all but one test site. This means that the probability of EHR detecting a ORAE in patients who had a true ORAE is close to 100%. Similarly, NPV is 100% in all but one test site. This suggests that the probability of EHR detecting a at-risk patient was also a patient at risk for ORAE based on the abstracted data is near perfect. Specificity is 100% in all test sites, indicating that the probability of correctly classifying a at-risk patient when the patient is truly and solely at risk for ORAE is 100%.

Overall, results from **Tables 11 through 14** suggest that the probability of EHR detecting a true ORAE in patients that indeed had an ORAE is nearly perfect, and that the measure has reasonably strong score level validity.

We will continue to evaluate measure validity through reevaluation as hospitals participate in this measure.

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

**2b2.EXCLUSIONS ANALYSIS**

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

The measure does not have denominator or numerator exclusions, although patients that only received naloxone within the OR suite are removed from numerator consideration. Ultimately, we aim to capture a broad cohort of patients who were administered an opioid medication during the hospitalization, and thus at risk for over-sedation using opioids. While these patients are at risk, they should not experience extreme respiratory depression or over-sedation to require naloxone because the vast majority of these events (excessive use of opioids) are preventable through proper dosing, monitoring, and following best practices.

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

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

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

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

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

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Clinical characteristics, including gender, age, race/ethnicity, reasons for hospitalization, clinical status when patients arrive at the hospital, or comorbidities can influence the risk of harm occurring during a hospitalization. Therefore, if hospitals care for patients with different degree of risk, then it may be important to account for such case mix in order to compare hospital performance.

However, ORAEs should be avoidable regardless of patient risk, particularly when the opioid was given after patients have arrived at the hospital. We consider the following criteria in determining whether or not risk adjustment or risk-stratification is warranted for this measure, if:

1. patients are at risk of the harm regardless of their demographic and clinical characteristics;

2. the majority of incidents of harm are linkable to care provision under the hospital control, for example harms caused by excessive or inappropriate medication dosing; and

3. there is evidence that the risk of harm can be largely reduced by following best care practices independent of patient inherent risks. For example, patients with multiple risk factors can still avoid the harm event when providers adhere to care guidelines.

In the case of ORAE, there is evidence that most instances of over-sedation requiring naloxone for reversal are avoidable. While certain patients may require higher doses to achieve pain control or are more sensitive to opioids (depending on their age, sex, and weight), the most common cause is hospital administration of excessive doses and inadequate monitoring. Because the dosing of opioids and the intensity of patient monitoring is entirely under the control of providers in hospitals, risk of ORAE can be reduced by following best practices. We thus do not think risk adjustment or risk-stratification is warranted for this measure.

To provide supportive evidence to our clinical rationale for not risk adjusting or risk-stratifying, we examined the measure performance rate in various subgroups of population. Of note, these summary statistics are derived from a small dataset that is by no means generalizable to the entire population. **Tables 15 and 16** present the results.

**Table 15. Summary Statistics of Measure Performance Rate (Sites 1-3)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Test Site 1** | | | **Test Site 2** | | | **Test Site 3** | | |
|  | **Mean** | **Std.Dev** | **P50** | **Mean** | **Std.Dev** | **P50** | **Mean** | **Std.Dev** | **P50** |
| Across all denominator patient-encounters | 0.11% | 3.30% | 0.00% | 0.34% | 5.78% | 0.00% | 0.45% | 6.68% | 0.00% |
| *Sub-groups* |  |  |  |  |  |  |  |  |  |
| Age bins |  |  |  |  |  |  |  |  |  |
| 18-35 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.35% | 5.90% | 0.00% |
| 36-64 | 0.16% | 3.98% | 0.00% | 0.42% | 6.45% | 0.00% | 0.73% | 8.50% | 0.00% |
| 65+ | 0.13% | 3.61% | 0.00% | 0.43% | 6.52% | 0.00% | 0.32% | 5.63% | 0.00% |
| Sex |  |  |  |  |  |  |  |  |  |
| Male | 0.30% | 5.51% | 0.00% | 0.59% | 7.64% | 0.00% | 0.15% | 3.88% | 0.00% |
| Female | 0.00% | 0.00% | 0.00% | 0.21% | 4.61% | 0.00% | 0.63% | 7.88% | 0.00% |
| Race |  |  |  |  |  |  |  |  |  |
| Black or African American | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| White | 0.11% | 3.35% | 0.00% | 0.35% | 5.93% | 0.00% | 0.47% | 6.81% | 0.00% |
| Other | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Unknown | – | – | – | 0.00% | – | 0.00% | 0.00% | 0.00% | 0.00% |
| Ethnicity |  |  |  |  |  |  |  |  |  |
| Hispanic or Latino | – | – | – | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Non-Hispanic | – | – | – | 0.35% | 5.90% | 0.00% | 0.45% | 6.70% | 0.00% |
| Unknown | 0.11% | 3.30% | 0.00% | 0.00% | 0.00% | 0.00% | – | – | – |
| (Primary) Payer |  |  |  |  |  |  |  |  |  |
| Medicare | 0.11% | 3.30% | 0.00% | 0.34% | 5.86% | 0.00% | 0.59% | 7.65% | 0.00% |
| Medicaid | 0.22% | 4.64% | 0.00% | 0.00% | 0.00% | 0.00% | 0.45% | 6.74% | 0.00% |
| Private Insurance | 0.00% | 0.00% | 0.00% | 0.30% | 5.45% | 0.00% | 0.20% | 4.51% | 0.00% |
| Self-pay or Uninsured | 0.00% | – | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Other | 0.00% | 0.00% | 0.00% | 3.03% | 17.27% | 0.00% | 0.00% | 0.00% | 0.00% |
| Unknown | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Table 16. Summary Statistics of Measure Performance Rate (Sites 4-6)**

|  | **Test Site 4** | | | **Test Site 5** | | | **Test Site 6** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Std.Dev** | **P50** | **Mean** | **Std.Dev** | **P50** | **Mean** | **Std.Dev** | **P50** |
| Across all denominator patient-encounters | 0.45% | 6.71% | 0.00% | 0.33% | 5.74% | 0.00% | 0.35% | 5.88% | 0.00% |
| *Sub-groups* |  |  |  |  |  |  |  |  |  |
| Age bins |  |  |  |  |  |  |  |  |  |
| 18-35 | 0.03% | 1.73% | 0.00% | 0.08% | 2.77% | 0.00% | 0.08% | 2.90% | 0.00% |
| 36-64 | 0.50% | 7.06% | 0.00% | 0.35% | 5.93% | 0.00% | 0.40% | 6.31% | 0.00% |
| 65+ | 0.79% | 8.87% | 0.00% | 0.44% | 6.60% | 0.00% | 0.49% | 7.00% | 0.00% |
| Sex |  |  |  |  |  |  |  |  |  |
| Male | 0.78% | 8.79% | 0.00% | 0.36% | 5.96% | 0.00% | 0.31% | 5.55% | 0.00% |
| Female | 0.27% | 5.14% | 0.00% | 0.31% | 5.58% | 0.00% | 0.38% | 6.13% | 0.00% |
| Race |  |  |  |  |  |  |  |  |  |
| Black or African American | 0.13% | 3.60% | 0.00% | 0.29% | 5.34% | 0.00% | 0.29% | 5.40% | 0.00% |
| White | 0.45% | 6.66% | 0.00% | 0.34% | 5.82% | 0.00% | 0.36% | 5.96% | 0.00% |
| Other | 0.54% | 7.31% | 0.00% | 0.51% | 7.15% | 0.00% | 0.39% | 6.25% | 0.00% |
| Unknown | 0.87% | 9.30% | 0.00% | 0.18% | 4.19% | 0.00% | 0.16% | 4.05% | 0.00% |
| Ethnicity |  |  |  |  |  |  |  |  |  |
| Hispanic or Latino | 0.45% | 6.67% | 0.00% | 0.25% | 5.01% | 0.00% | 0.22% | 4.72% | 0.00% |
| Non-Hispanic | 0.45% | 6.66% | 0.00% | 0.35% | 5.92% | 0.00% | 0.40% | 6.35% | 0.00% |
| Unknown | 0.70% | 8.38% | 0.00% | 0.29% | 5.36% | 0.00% | 0.29% | 5.35% | 0.00% |
| (Primary) Payer |  |  |  |  |  |  |  |  |  |
| Medicare | 0.79% | 8.83% | 0.00% | 0.43% | 6.52% | 0.00% | 0.54% | 7.36% | 0.00% |
| Medicaid | 0.34% | 5.80% | 0.00% | 0.36% | 5.97% | 0.00% | 0.20% | 4.43% | 0.00% |
| Private Insurance | 0.22% | 4.68% | 0.00% | 0.29% | 5.34% | 0.00% | 0.28% | 5.31% | 0.00% |
| Self-pay or Uninsured | 0.49% | 6.96% | 0.00% | 0.13% | 3.58% | 0.00% | 0.00% | 0.00% | 0.00% |
| Other | 0.42% | 6.45% | 0.00% | 0.00% | 0.00% | 0.00% | 0.41% | 6.37% | 0.00% |
| Unknown | 0.00% | 0.00% | 0.00% | – | – | – | – | – | – |

**Tables 15 and 16** reveal three points that are worth noting. First, measure performance rates ranged from 0.11% (for every 1,000 qualified hospital admissions there are 1.1 inpatient encounters where patients suffered ORAE) to 0.45% (for every 1,000 qualified hospital admissions there are 4.5 inpatient encounters where patients suffered ORAE), indicating ample room for quality improvement in hospital inpatient environment. Second, larger hospitals (e.g., test sites 4 to 6), though having more qualified admissions (**Tables 2 and 3**), do not necessarily have higher rates of ORAE. This suggests that all hospitals, irrespective of size, need to follow best practices in patient care to prevent the incidence of ORAE. Third, in four of the six test sites, male patients were showing higher likelihoods of experiencing ORAEs even though female patients were more likely to be at risk (**Tables 2 and 3**). Patients who are White were more likely to be at risk (**Tables 2 and 3**), and yet do not have consistently higher odds of experiencing ORAEs. Elderly patients (age 65 or older) tended to experience ORAEs more often than patients who were younger, but the difference in magnitude is modest. Overall, **Tables 15 and 16** show no clear pattern in measure performance rates across subgroups of population. These provide supportive evidence to the clinical rationale we provided above.

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

**2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

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

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

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

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

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

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

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

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

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

We examined data to determine if there were meaningful differences in measure performance rates across the six implementation test sites. In particular, we calculated the confidence intervals around the performance rate estimates and the variation in measure performance rates among test sites 1 through 6. We used these statistics to determine the stability of the rate estimate and if there were differences in measure performance rates between sites, respectively.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Across test sites 1 through 6, the average measure performance rate was 0.36% with the 95% confidence interval equal to 0.31% and 0.41%. The fairly narrow confidence interval suggests that the measure performance rate was estimated with precision.

However, across the six test sites the measure performance rate ranged from 0.11% to 0.45%. The relatively wide variability suggests that there exists ample room for quality improvement.

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Results from **Tables 15 and 16** showed that the rate of ORAE estimated using the patient EHR data from calendar year 2019 were within the range of harm rates found in the literature, which was between 0.1% and 1.3% among studies using naloxone administration as a surrogate measure of respiratory depression (Cashman, 2004). The relatively wide variability in the rate of ORAE across the six sites demonstrates that there exists room for improvement in reducing the ORAE among at-risk patients.

Reference:

Cashman, J. N., and S. J. Dolin. "Respiratory and haemodynamic effects of acute postoperative pain management: evidence from published data." *British Journal of Anaesthesia* 93, no. 2 (2004): 212-223.

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

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

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

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

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

As mentioned in section 2a2.2, we quantitatively assessed data element reliability using the rate of missing or erroneous data for every critical data element needed for measure implementation.

For the critical data elements used in the measure, we anticipate that there should be no missing data and, if any, the rate would approximate zero. This is because the measure uses variables that are expected to be available in structured fields of the EHR and captured as part of the routine care.

**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*)

**Tables 17 and 18** (reprinted from the above) display the percentage of missing or erroneous data for the critical data elements needed for measure implementation. The results suggest that all critical data elements are reliably and consistently captured in patient EHRs.

**Table 17. Data Element Reliability Results (Frequency of Missing or Erroneous Data) for the Critical Data Elements (Sites 1-3)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Test Site 1** | | | **Test Site 2** | | | **Test Site 3** | | |
| Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) |
| Patient inpatient encounter discharge DateTime | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Patient birth date | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Opioid administration DateTime | 0 | 1,839 | 0% | 0 | 2,089 | 0% | 0 | 1,784 | 0% |
| Naloxone Administration DateTime | 0 | 2 | 0% | 0 | 7 | 0% | 0 | 8 | 0% |
| Naloxone Administration Location | 0 | 2 | 0% | 0 | 7 | 0% | 0 | 8 | 0% |

**Table 18. Data Element Reliability Results (Frequency of Missing or Erroneous Data) for the Critical Data Elements (Sites 4-6)**

| **Data Element** | **Test Site 4** | | | **Test Site 5** | | | **Test Site 6** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) | Missing (or error) Count (#) | Patient Encounters (#) | Missing (or error) Percent (%) |
| Patient inpatient encounter discharge DateTime | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Patient birth date | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Opioid administration DateTime | 0 | 11,280 | 0% | 0 | 13,310 | 0% | 0 | 18,435 | 0% |
| Naloxone Administration DateTime | 0 | 51 | 0% | 0 | 44 | 0% | 0 | 64 | 0% |
| Naloxone Administration Location | 0 | 51 | 0% | 0 | 44 | 0% | 0 | 64 | 0% |

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*).

For all the critical data elements required for the measure implementation, we found no missing or erroneous data for all the six implementation test sites. The results suggest that all critical data elements are reliably and consistently captured in patient EHRs, and that measure implementation is feasible.