NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation.

Brief Measure Information

NQF #: 3501e

Corresponding Measures:

De.2. Measure Title: Hospital Harm – Opioid-Related Adverse Events

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: This measure assesses the proportion of inpatient hospital encounters where patients ages 18 years of age or older have been administered an opioid medication, subsequently suffer the harm of an opioid-related adverse event, and are administered an opioid antagonist (naloxone) within 12 hours. This measure excludes opioid antagonist (naloxone) administration occurring in the operating room setting.

1b.1. Developer Rationale: Opioids are often the foundation for sedation and pain relief. However, use of opioids can also lead to serious adverse events, including constipation, over sedation, delirium, and respiratory depression. Opioid-related adverse events have both patient-level and financial implications. Patients who experience this event have been noted to have 55% longer lengths of stay, 47% higher costs, 36% higher risk of 30-day readmission, and 3.4 times higher payments than patients without these adverse events (Kessler et al., 2013).

Most opioid-related adverse events are preventable. Of the adverse drug events reported to the Joint Commission´s Sentinel Event database, 47% were due to a wrong medication dose, 29% to improper monitoring, and 11% to other causes (e.g., medication interactions, drug reactions) (Joint Commission, 2012; Overdyk, 2009). Additionally, in a closed-claims analysis, 97% of adverse events were judged preventable with better monitoring and response (Lee et al., 2015) . Naloxone administration is often used as an indicator of a severe opioid-related adverse event, and implementation of this measure can advance safe use of opioids in hospitals and prevent these serious and potentially lethal adverse drug events.

Naloxone is an opioid reversal agent typically used for severe opioid-related adverse events. Naloxone administration has been used in a number of studies as an indicator of opioid-related adverse events (Nwulu et al., 2013; Eckstrand et al., 2009).

From Part 10 of the 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care (Lavonas et al., 2015), the following recommendation is listed for use of Naloxone:

Naloxone is a potent opioid receptor antagonist in the brain, spinal cord, and gastrointestinal system. Naloxone has an excellent safety profile and can rapidly reverse central nervous system (CNS) and respiratory depression in a patient with an opioid-associated resuscitative emergency.

References:

Eckstrand, J. A., Habib, A. S., Williamson, A., Horvath, M. M., Gattis, K. G., Cozart, H., & Ferranti, J. Computerized surveillance of opioid-related adverse drug events in perioperative care: a cross-sectional study. Patient Saf Surg. 2009;3(1), 18.

Kessler ER, Shah M, Gruschkus SK, Raju A. Cost and quality implications of opioid-based postsurgical pain control using administrative claims data from a large health system: opioid-related adverse events and their impact on clinical and economic outcomes. Pharmacotherapy. 2013;33(4):383-391.

Lavonas EJ, Drennan IR, Gabrielli A, Heffner AC, Hoyte CO, Orkin AM, Sawyer KN, Donnino MW. Part 10: Special Circumstances of Resuscitation: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2015 Nov 3;132(18 Suppl 2):S501-18. doi: 10.1161/CIR.0000000000000264. Erratum in: Circulation. 2016 Aug 30;134(9):e122.

Lee, L. A., Caplan, R. A., Stephens, L. S., Posner, K. L., Terman, G. W., Voepel-Lewis, T., & Domino, K. B. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology. 2015:122(3), 659-665.

Nwulu, U., Nirantharakumar, K., Odesanya, R., McDowell, S. E., & Coleman, J. J. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. Eur J Clin Pharmacol. 2013;69(2), 255-259.

Overdyk FJ: Postoperative respiratory depression and opioids. Initiatives in Safe Patient Care, Saxe Healthcare Communications, 2009

The Joint Commission. Safe use of opioids in hospitals. Sentinel Event Alert. 2012(49):1-5. https://www.jointcommission.org/-/media/deprecated-unorganized/imported-assets/tjc/system-folders/topics-library/sea_49_opioids_8_2_12_finalpdf.pdf?db=web&hash=0135F306FCB10D919CF7572ECCC65C84

S.4. Numerator Statement: Inpatient hospitalizations where an opioid antagonist (naloxone) was administered outside of the operating room and within 12 hours following administration of an opioid medication. Only one numerator event is counted per encounter.

S.6. Denominator Statement: Inpatient hospitalizations for patients 18 years or older during which at least one opioid medication was administered. An inpatient hospitalization includes time spent in the emergency department or in observation status when the patients are ultimately admitted to inpatient status.

S.8. Denominator Exclusions: N/A; there are no denominator exclusions

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Records

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- The logic model presented by the developer for this outcome measure links evidence-based practices such as routine patient monitoring for potential adverse effects of opioids, proper dosing of opioids and proper drug selection with better quality of care associated with excessive opioid administration in the hospital setting.
- The goal of this eCQM as stated by the developer, is to improve safety for patients who receive opioids during their hospitalization.
- The developer also cited literature highlighting the variability in hospital monitoring practices, which suggests opportunities for improvement. Additionally, the developer cited evidence that implementing targeted interventions can have a significant impact on opioid-related adverse events. These include enhanced monitoring practices, improved clinical decision support in the electronic medical record (EMR), and various adjustments to dosing for high-risk patients that included clinician education.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

Does the measure assess performance on a health outcome (Box 1) -> (yes) -> Is there a relationship between the measure and at least one healthcare action is demonstrated by empirical data (Box 2) -> (yes) -> PASS

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- This eCQM was tested in six hospitals across five different states with varying bed sizes (between 71 to over 500 beds). Half of the test sites utilized Meditech and the other half, Cerner as the EHR vendor.
- Data from all six testing sites were collected between 1/1/2019 - 12/31/2019.
- The range in performance across tested hospitals was from 0.11%-0.45%.
- The overall performance rate for the six hospitals was 0.34% with a standard deviation of 0.12%.
- 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.

Disparities

- The developer acknowledges that summary statistics are calculated at the encounter-level and derived from a sample of six hospitals and may not be generalizable to the entire population.
- The measure performance was stratified for disparities by age, sex, race, ethnicity, and payer source for each of the six Beta Implementation Test Sites.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures–are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- Evidence appears appropriate
- Evidence is strong and supported by practice recommendations and empirical data
- Opioids are often used for sedation and pain relief. But when used improperly, these drugs can cause serious adverse reactions and patient harm. Opioid-related adverse events are often preventable with better monitoring and response. Naloxone is an opioid antagonist used to rapidly reverse opioid overdose. The developer argued that its use can be "an indicator of severe opioid-related adverse events, and implementation of this measure can advance the safe use of opioids in hospitals and prevent these serious and potentially lethal adverse drug events." So the evidence is strong in supporting this outcome measure.
- Use of naloxone correlated with adverse events as a result of opioid prescribing.
- Evidence is Pass.
- Change? Compile doses and symptoms that elicited naloxone use.
- moderate level evidence
- NA
- outcome measure

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- There is a performance gap noted. Measure performance was stratified by age, gender, ethnicity, race and payer. would like to hear more about these findings.
- While variability present, the absolute numbers are small and thus single events may skew performance data. Do the measure developers have thoughts on the robustness of this measure to very small numbers of events?
- This measure was tested in six hospitals across five different states. The performance score ranges from 0.11% to 0.45%, demonstrating a moderate gap in care to justify this measure. The measure performance was stratified for disparities by age, sex, race, ethnicity, but no risk adjustment or risk stratification is applied to this measure.
- Adverse events in hospitals can be avoided by monitoring this metric and improving the practice of opioid prescribing. Indirectly measuring use of naloxone as an indicator of adverse opioid-related events is a reasonable way to improve patient quality of care.
- I feel the reported range of 0.11-0.45% is a VERY small range, and I don't believe that's sufficient variability to warrant a national performance measure. My vote on performance gap is "LOW"
- Gap exists. Performance measure justified.
- moderate

- NA
- moderate gap

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

NQF eCQM Evaluation Summary

- Submitted measure specification follows established technical specifications for eCQMs (HQMF, QDM, and CQL) as indicated Sub-criterion 2a1.
- Submitted measure specification is fully represented and is not hindered by any limitations in the established technical specifications for eCQMs.
- The Feasibility Scorecard indicated the *Encounter, performed: Encounter Inpatient Facility Location: Operating Room Suite* data element has feasibility issues related to accuracy (see Feasibility section below). Take this into account while reviewing at the validity testing for this data element and the measure.
- The following is the developer's plan for addressing the accuracy/feasibility issue for this data element:
    - *"~8% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Workflow modifications would better enable capture (e.g., use of EHRs modules already available through vendor system to track temporary locations)"*

2b2-2b6. Potential threats to validity should be assessed/addressed.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel and discussed on the call. A summary of the measure and the Panel discussion is provided below.

Reliability

- For data element reliability, the developer compared electronically extracted data to manually abstracted data using kappa to quantify agreement. The Kappa coefficient was 0.98 at one site and 1.00 at all other sites for the six randomly selected sub-samples, comparing the electronically extracted EHR data to manually extracted EHR data for the same medical record.
- There were concerns by some SMP members that measure score reliability was not conducted.

Validity

- Data element: Tested inter-rater agreement by comparing the hospitals' EHR data to a clinical abstractor (as described above in the reliability section). The agreement rate between data electronically extracted from the sampled patients' EHR and data manually abstracted from the medical records was 100 percent for all but two data elements. Measure score validity was assessed for this rather small sample by PPV, sensitivity, NPV, and specificity. PPV was 100 percent, and sensitivity is 100 percent in all but one test site. NPV is also 100 percent. Specificity is 100 percent.
- Score-level: An EHR-reported opioid related adverse event was compared to clinical review of the patient's chart.
- There were concerns by some SMP members about the validity testing approach. In addition, there were some concerns that the measure was not risk-adjusted.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity:	☐ High	☒ Moderate	☐ Low	☐ Insufficient

Committee Pre-evaluation Comments:

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- My concerns are about the workflow changes that would be needed based on the experience of some of the beta facilities. In hospitals changes to workflow can take many steps. How will this be supported in hospitals?
- No reliability concerns as kappa statistics were high. However, samples were small as were the number of hospitals; I would guess as heterogeneity of sites increases this will warrant further evaluation
- One SMP member raised a valid question on "how the measure can differentiate between the use of Naloxone as an indicator of opioid-related adverse events vs. other uses following or in combination with opioid use." I would be interesting to see clarification by the developer. The developer stated that "~8% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Workflow modifications would better enable capture (e.g., use of EHRs modules already available through vendor system to track temporary locations)." I would be interested in learning more from the developer on how this measure can be implemented consistently.
- This is a straightforward measure.
- Agree with the methods panel's assessment.
- No concerns
- moderate
- NA
- panel review - measure score reliability not conducted

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure; reliability testing and results for the measure?

- see above
- No
- When comparing between data extracted from the EHR electronically and data extracted from patient medical records, the Kappa coefficient was 0.98 at one site and 1.00 at all other sites for the six randomly selected sub-samples. This indicates an excellent agreement. I would rate the reliability as high.
- no
- No concerns
- none
- no concerns
- NA
- panel rated moderate

2b1. Validity -Testing: Do you have any concerns with the validity testing and results for the measure?

- no concerns
- Yes. Validity testing was a very small number of sites and using EHR vs clinical records is limited. I think comparisons to other quality measures may be appropriate, though with a small number of sites these comparisons may have substantial noise.

- No concerns. The validity is similar as reliability testing. The preliminary rating is moderate.
- No concerns
- Agree with the methods panel's assessment.
- no
- no concerns
- NA
- no concerns - rated moderate

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- Again, just the workflow issues raised previously.
- These may be records collected outside EHR or in paper notes.
- This measure was tested in six hospitals across five different states. The performance score ranges from 0.11% to 0.45%. This shows a meaning difference in care quality. No issues with missing data. The developer does not anticipate any missing data. If there are, "the rate would be approximate zero because the measure uses variables that are expected to be available in structured fields of the EHR and captured as a part of the routine care."
- no concerns
- No concerns.
- no.
- no concerns
- NA
- no concerns

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- no risk adjustment
- Risk adjustment may be needed, as certain conditions may increase risk. Why no exclusions for patients on hospice, cancer, sickle cell?..I understand appropriate prescribing needs to take this into account but I wonder if certain hospitals who care for these populations disproportionately may have higher rates? Also, why limit to those using opioids at least once in hospital? Why not have a rate per 100 patients or something to help protect hospitals which appropriate limit opioid use but perhaps have a small number of events skewing their results?
- No exclusions and risk adjustment for this measure, which I think is reasonable. Regardless of a patient's risk, age, sex, or socioeconomic, no patient should be subject to risk of improper use of opioids and medication harm and the resulting adverse events should be avoidable.
- Measure was not risk adjusted.
- I think for this measure no risk adjustment is appropriate.

- unnecessary
- no concerns
- NA
- no concerns

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Of all sites used for the measure feasibility assessment, some reported that their anesthesiologists document their activities on paper-based anesthesia records inside of the OR rather than via the electronic medication administration record (eMAR). This suggests that, at this time, for these sites, opioid and naloxone administration inside of the OR will not be available for structured electronic extraction or appear in patient EHRs.
- For opioid and naloxone administration outside of OR suite, however, all test sites confirmed that they are documented in the eMARs, and available for electronic extraction.
- The scorecard indicated the following data elements have feasibility issues due to either availability or standards indicating that the data element may not be available electronically or have a credible near-term path to electronic collection.
    - Encounter, performed: Encounter Inpatient Facility Location: Operating Room Suite
        - developer's plan for addressing the feasibility issue for this data element: *"~8% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Workflow modifications would better enable capture (e.g., use of EHRs modules already available through vendor system to track temporary locations)"*
    - Medication Administered: Opioids, All
    - Medication, Administered: Opioid Antagonist
        - developer's plan for addressing the feasibility issue for these data elements:
        - *"~22% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Clinical and technical workflow modifications would better enable capture (i.e.., use of anesthesia modules already available through the vendor systems--Cerner and Allscripts)"*

Questions for the Committee:

- Is the data collection strategy ready to be put into operational use?
- Does the eCQM Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- as stated previously, the workflow issues raised may require more attention
- Extractions from EMR data, but it seems some modules and hospitals use paper based records which may make initial adoption less feasible.
- The developer reported a few feasibility issues. For instance, "of all sites used for the measure feasibility assessment, some reported that their anesthesiologists document their activities on paper-based anesthesia records inside of the OR rather than via the electronic medication administration record (eMAR). This suggests that, at this time, for these sites, opioid and naloxone administration inside of the OR will not be available for structured electronic extraction or appear in patient EHRs." However, for opioid and naloxone administration outside of OR suite, all test sites confirmed that they are documented in the eMARs, and available for electronic extraction. Given that non-anesthesia-related opioid administrations are already captured electronically, the developer is optimistic that measure implementation is still feasible and believes measure implementation will drive workflow changes toward electronic capture within the OR. Feasibility is preliminarily rated as moderate.
- Feasibility issues were reported by about 22% of the facilities indicating that there would be difficulties obtaining the data required.
- This seems low as up to 22% of sites may not have easy access to needed data elements.
- none.
- moderate rating
- NA
- moderate feasibility

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The measure is not currently in use in any accountability programs.
- Following MAP 2021-2022 review, the developer envisions that this measure will be considered for accountability programs via future rulemaking.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The measure is currently not implemented in a public reporting or accountability program.
- Implementation is planned pending finalization of the NQF endorsement and CMS rulemaking processes.

Additional Feedback: N/a

***Questions for the Committee**:*

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- This eCQM is not currently used in any quality improvement program, but a primary goal of the eCQM is to provide hospitals with performance information necessary to implement focused quality improvement efforts.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer did not identify any unintended consequences during eCQM development or testing.

- To verify that the eCQM , as currently specified, does not detect false positives, the developer conducted empirical tests to examine whether numerator cases identified by the measure are true positives.

Potential harms

- There are no harms identified by the developer.

Additional Feedback: N/a

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- not publicly reported, not used in accountability programs yet
- Measure feedback provided by stakeholder groups in a TEP, and feedback was implemented to tailor measure more specifically.
- This outcome measure is currently not used in any accountability and public reporting programs. Following MAP 2021-2022 review, the developer envisions that this measure will be considered for accountability programs via future rulemaking. Final measure specifications for implementation will be made publicly available on CMS' appropriate quality reporting website, once finalized through the NQF endorsement and CMS rulemaking processes.
- The measure is not currently publicly reported and not used in any accountability programs.
- No concerns
- Limited feedback identified
- no concerns
- NA
- no real current use, plan for accountability

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- no unintended consequences evident
- No reports of unintended consequences. Benefits likely outweigh harms but small % of patients experiencing outcome make this less certain
- While this measure is not currently used in any quality improvement program, the developer stated that a primary goal of the eCQM is to provide hospitals with performance information necessary to implement focused quality improvement efforts. No unintended consequences were identified by the developer. The preliminary rating for usability is moderate.
- This eCQM is not currently used in any quality improvement program but the primary goal is to provide performance information necessary to implement focused quality improvement efforts.
- No concerns
- Unclear until placed into widespread use. Clearly beneficial, but there may be a tendency to not use naloxone when it would benefit the patient slightly overdosed on an opioid.
- benefits > harms
- NA

- moderate usability

Criterion 5: Related and Competing Measures

Related or competing measures

- 3316e: Safe Use of Opioids – Concurrent Prescribing
- 3389: Concurrent Use of Opioids and Benzodiazepines (COB)

Harmonization

- As a result of the varying measure focuses, the Hospital Harm – Opioid Related Adverse Events eCQM has a broad denominator of all inpatient adults >=18 years who received a hospital administered opioid, while NQF #3316e has a narrower denominator of adults >=18 years prescribed an opioid or benzodiazepine at discharge from a hospital-based encounter.
- NQF #3316e excludes patients with an active cancer diagnoses, palliative care order, or length of stay >120 days.
- NQF #3389 addresses outpatient prescription claims and excludes patients in hospice, or with a cancer or sickle cell disease diagnosis.

Committee Pre-evaluation Comments: Criterion 5:

Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- no competing measures
- Overlaps are appropriate, given different settings and populations
- Two other similar opioid-related measures are identified, NQF #3316e and #3389. But they address a different patient population or concurrent use of opioid and Benzo. So I do not think they are competing measures.
- unknown
- No concerns.
- no
- documented in worksheet
- NA
- 2 competing measures with differing populations and exclusions

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/03/2021

- No NQF Members have submitted support/non-support choices as of this date.
- No Public or NQF Member comments submitted as of this date.

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3501e

Measure Title: Hospital Harm - Opioid-Related Adverse Events

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☒ Yes ☒ No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   - Submitted measure specification follows established technical specifications for eCQMs (HQMF, QDM, and CQL) as indicated Sub-criterion 2a1
   - Submitted measure specification is fully represented and is not hindered by any limitations in the established technical specifications for eCQMs

2. **Briefly summarize any concerns about the measure specifications.**

   Panel Member 1: No concerns.

   Panel Member 3: It may be good to clarify how the measure can differentiate between use of Naloxone as an indicator of opioid-related adverse events vs. other uses following or in combination with opioid use, to strengthen the rationale of the measure. Is there no way to directly identify an opioid-related adverse event from the EHR without using Naloxone as the signal of this measure? I am sure this has been thought out, but an explanation would strengthen this submission.

   Panel Member 4: No concerns.

   Panel Member 6: None

   Panel Member 7: Why OR only? Why not PACU? Or ICU? How can this be measured? Is oral naloxone included or IV only? Nursing notes - meaning?

   Panel Member 8: None

RELIABILITY: TESTING

Type of measure:

☒ Outcome (including PRO-PM) ☐ Intermediate Clinical Outcome ☐ Process

☐ Structure ☐ Composite ☐ Cost/Resource Use ☐ Efficiency

Data Source:

☐ Abstracted from Paper Records ☐ Claims ☐ Registry
☒ Abstracted from Electronic Health Record (EHR) ☒ eMeasure (HQMF) implemented in EHRs
☐ Instrument-Based Data ☐ Enrollment Data ☐ Other (please specify)

Level of Analysis:

☐ **Individual Clinician** ☐ **Group/Practice** ☒ **Hospital/Facility/Agency** ☐ **Health Plan**
☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**
☐ **Integrated Delivery System** ☐ **Other (please specify)**

Measure is:

☐ **New** ☒ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☒ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing of patient-level data** conducted?

   ☒ **Yes** ☐ **No**
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Panel Member 1: Tested inter-rater agreement by comparing the hospitals' EHR data to a clinical abstractor.

   Panel Member 2: The developer established data element reliability by comparing electronically extracted data to manually abstracted data, using kappa to quantify agreement.

   Panel Member 3: Developers gave a detailed explanation on why the available data for testing would not allow an accurate representation of a score level reliability analysis.

   Panel Member 4: Appropriate for data element. No/minimal performance score testing due to inadequate number of hospitals participating in the testing.

   Panel Member 5: Data Element Reliability Results (Cohen's Kappa) for the Critical Data Elements was high to perfect agreement

   Panel Member 6: Two methods were utilized to estimate reliability of the data elements. First, the rate of missing or erroneous data was calculated. Secondly, for six sub-samples of 100 patient encounters randomly selected, a kappa statistic for inter-rater reliability was calculated. Signal to noise methodology would be utilized for larger sample size, but this sample size was not felt to be adequate.

   Panel Member 7: Data: Rate of wrong/missing + Cohen's kappa Score: beta binomial was problematic and "not performed"

   Panel Member 8: Data element testing using kappa (claims to chart review). This appears to be a test of validity rather than reliability. I think this is OK because a test of data element validity for a claims/EHR based measure counts as reliability evidence. Given the extremely low event rates <1%, its unclear if the reliability testing included any events in the random sample of cases. Did not perform measure score reliability assessment.
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

Panel Member 1: 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").

Panel Member 2: The results indicated perfect agreement.

Panel Member 3: I have no concerns with the data element reliability analysis which demonstrated excellent results.

Panel Member 4: High level of concordance with EMR and abstracted values of the data elements.

Panel Member 5: Critical data elements were found to be very reliable. Score level reliability would need further testing with more data as noted

Panel Member 6: Kappa was 0.98 at one site and 1.00 at all other sites for the six randomly selected sub-samples, comparing the electronically extracted EHR data to manually extracted EHR data for the same medical record.

Panel Member 7: Data element validity: "all but two data elements showed a match rate of 100%, indicating that valid and accurate data were extracted from patient EHRs." Not clear how "erroneous" was defined? Extreme? Compared to a gold standard? Error rates are 0% and Kappas are 0.98. I would like to understand better how this is plausible. I would like to understand how nursing notes are "validated."

Panel Member 8: Perfect agreement on the data element level.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☒ **No**

   ☒ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☒ **No**

   ☐ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and **all** testing results):

    ☒ **High** (NOTE: Can be HIGH **only if** score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has **not** been conducted)

    ☐ **Low** (NOTE: Should rate **LOW** if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☒ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

Panel Member 1: Used appropriate methods to test the reliability of the critical data elements. Found high level of agreement in data values taken from the EHR as compared to a clinical abstractor.

Panel Member 2: Only data element reliability was empirically tested. Simulation results on measure score reliability were informative but could not be used as evidence.

Panel Member 3: I have no concerns with the data element reliability analysis which demonstrated excellent results.

Panel Member 4: Based on data elements testing results and minimal/no performance score testing.

Panel Member 5: As mentioned above, critical data elements very high agreement, measure score would need further assessment when more data is collected.

Panel Member 6: Kappa was .98 at one location and 1.00 at the other five locations.

Panel Member 7: Noble effort. Additional scrutiny of data element validity would be valuable. No comment on measure reliability (SNR).

Panel Member 8: The data element validity testing presented in this section is acceptable evidence of reliability at that level.

VALIDITY: TESTING

12. **Validity testing level:** ☐ **Measure score** ☒ **Data element** ☒ **Both**
13. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

    Submission document: *Testing attachment, section 2b1.*

    ☒ **Yes**

    ☒ **No**

    ☐ **Not applicable** (data element testing was not performed)
14. **Method of establishing validity of the measure score:**

    ☐ **Face validity**

    ☒ **Empirical validity testing of the measure score**

    ☒ **N/A (score-level testing not conducted)**
15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☒ **No**

    ☒ **Not applicable** (score-level testing was not performed)
16. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    Panel Member 1: Data element: Tested inter-rater agreement by comparing the hospitals' EHR data to a clinical abstractor. Score-level: Compared a EHR-reported ORAE to clinical review of the

patient's chart, which given the strict definition of the numerator event does not appear to be much different than validating the data elements that comprise the measure.

Panel Member 2: What is described under measure score validity is really about data elements (numerator and denominator) validity, so it is still at data element level, not score level. Indices such as PPV, sensitivity, NPV, and specificity are all pertaining to defining numerator and denominator, not measure score.

Panel Member 3: What is described under measure score validity is really about data elements (numerator and denominator) validity, so it is still at data element level, not score level. Indices such as PPV, sensitivity, NPV, and specificity are all pertaining to defining numerator and denominator, not measure score.

Panel Member 4: As this is an eCQM, the developer assessed the accuracy of the data elements through abstraction of the medical record. They used PPV and NPV to determine the percent accuracy between the EMR data and the abstracted data. The developer determined that the accuracy was fairly high for both PPV and NPV.

Panel Member 5: 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

Panel Member 6: Data element validity: Agreement rate between data electronically extracted from the sampled patients' EHR and data manually abstracted from the medical records was 100% for all but two data elements Measure score validity was assessed for this rather small sample by PPV, sensitivity, NPV, and specificity. PPV was 100%, sensitivity is 100% in all but one test site. NPV is 100%. Specificity is 100%.

Panel Member 7: "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 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."

Panel Member 8: The data element validity test mirrors the methods in the reliability test, comparing electronically extracted values to expert chart review. The measure score validity testing appears to be at the patient-level, not the entity level: "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." They again found perfect agreement between e-extracted values and chart review.

17. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    Panel Member 1: Data element: 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"). Score-level: Found high levels of agreement between the EHR capture of a ORAE and clinical review.

Panel Member 2: All results combined supported data element validity.

Panel Member 3: No concerns

Panel Member 4: ? This is classified as a never event according to the developer, and given the degree of accuracy of the event being present or not, I believe their methods meet the criteria.

Panel Member 5: Yes, I do agree but recommend if implemented in a wider population for reviewing the validity when data from more sites available

Panel Member 6: In this small sample size, NPV, PPV, sensitivity and specificity were used to assess validity. In a larger sample size, other methodologies will be required.

Panel Member 7: There is no gold standard. I have concerns that these are measuring themselves or measuring error. This measure may be achievable but further scrutiny of data validity may be warranted.

Panel Member 8: They again found perfect agreement between e-extracted values and chart review.

Staff: The scorecard indicated the *Encounter, performed: Encounter Inpatient Facility Location: Operating Room Suite* data element has feasibility issues related to accuracy. Take this into account while reviewing at the validity testing for this data element and the measure.

The following is the developer's plan for addressing the accuracy/feasibility issue for this data element: "*~8% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Workflow modifications would better enable capture (e.g., use of EHRs modules already available through vendor system to track temporary locations)*"

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Panel Member 1: No exclusions for the measure.

Panel Member 3: There are no exclusions to this measure

Panel Member 4: No concerns

Panel Member 5: N/A

Panel Member 6: None

19. **Risk Adjustment**

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒ Yes ☒ No ☐ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☒ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☐ Yes ☒ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☒ No

19d. **Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☒ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

19d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: The developers suggest the reason to not risk-adjust is ORAEs should be avoidable regardless of patient risk.

Panel Member 2: The developer conceptual justification and empirical results to support why risk adjustment is not warranted for this measure.

Panel Member 3: I agree with the developers' rationale to not risk-adjust this measure, since an ORAE should be avoidable regardless of patient risk, with supporting evidence that it can be avoided regardless of patient risk.

Panel Member 4: Justification provided and no evidence that contradicts the developer's rationale

Panel Member 5: Justification for not risk adjusting acceptable

Panel Member 6: Risk adjusted not need for this small EHR validation study.

Panel Member 7: No risk adjustment: "ORAEs should be avoidable regardless of patient risk, particularly when the opioid was given after patients have arrived at the hospital." I do not share this view. 1. There is drug diversion in the hospital. 2. Some patients have specific conditions (or undergo specific procedures) making safe opioid administration very difficult (higher risk for even safe systems to cause harm).

Panel Member 8: I could not find the rationale for not risk adjusting this outcome measure.

20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

Submission document: Testing attachment, section 2b4.

Panel Member 1: No concerns. Performance rates ranged from 0.11% to 0.45%.

Panel Member 2: Given that the event rate is rare, it is important to quantify the precision of the rate when reporting. For small volume hospitals, precision may be an issue.

Panel Member 3: The analyses provided should be considered preliminary due to the low number of sites included. Across the six test sites the measure performance rate ranged from 0.11% to 0.45%. Although these rates are overall low and may seem at first to not demonstrate variability in performance, due to the gravity of an ORAE event, I am in agreement with the developers that this variation is evidence supporting there is enough room for improvement to justify the measure. However, the part that is missing, and should be completed when this measure comes back for

maintenance, are the site level reliability and validity analyses, to confirm that the measure is able to reliably and validly identify difference in performance between hospitals.

Panel Member 4: No concerns

Panel Member 5: No concerns but recommend testing at a later date after more hospitals begin reporting

Panel Member 6: Not in this preliminary measure as small sample size.

Panel Member 8: Across the six test sites the measure performance rate ranged from 0.11% to 0.45%. Whether any of the sites were statistically different from the average or each other was not shown.

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

    Submission document: Testing attachment, section 2b5.

    Panel Member 1: N/A

    Panel Member 3: NA

    Panel Member 5: None

    Panel Member 6: Not relevant

    Panel Member 7: Nursing notes are used to gauge the numerator. These are not standardized across hospitals…

22. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    Panel Member 1: Developers identified no missing data.

    Panel Member 3: No concerns

    Panel Member 4: No concerns

    Panel Member 5: None

    Panel Member 6: None

For cost/resource use measures ONLY:

23. **Are the specifications in alignment with the stated measure intent?**

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☒ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

    ☒ **Low** (NOTE: Should rate LOW if you believe that there **are** threats to validity and/or relevant threats to validity were **not assessed OR** if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level **is required**; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    Panel Member 1: The developers used an odd method for assessing score-level validity, but did test each critical data element. The data element testing produced strong results for each data element.

    Panel Member 3: See comments above

    Panel Member 4: Based on the information provided about testing and the characterization of the measure as a patient safety event.

    Panel Member 5: Test results for all participating sites were strong

    Panel Member 6: Small sample size but near 100% results.

    Panel Member 7: Good idea - a work in progress. Assumptions about data validity may require more robust validation. Risk adjustment may be required.

    Panel Member 8: I believe the data elements can be accurately extracted from the EHR. I am concerned about lack of score variability and lack of rationale for not risk adjusting the measure.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    Panel Member 1: None.

    Panel Member 6: None

    Staff: The scorecard indicated the following data elements have feasibility issues in due to either availability or standards indicating that the data element may not be available electronically or have a credible near-term path to electronic collection.

    - Encounter, performed: Encounter Inpatient Facility Location: Operating Room Suite
        - developer's plan for addressing the feasibility issue for this data element: "*~8% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Workflow modifications would better*

enable capture (e.g., use of EHRs modules already available through vendor system to track temporary locations)"

- Medication Administered: Opioids, All
- Medication, Administered: Opioid Antagonist
    - developer's plan for addressing the feasibility issue for these data elements:
    - *"~22% of the facilities used for feasibility assessment identified that there would be difficulties obtaining the data required for this element. Clinical and technical workflow modifications would better enable capture (i.e.., use of anesthesia modules already available through the vendor systems--Cerner and Allscripts)"*

Developer Submission

NQF #: 3501e

Corresponding Measures:

De.2. Measure Title: Hospital Harm – Opioid-Related Adverse Events

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: This measure assesses the proportion of inpatient hospital encounters where patients ages 18 years of age or older have been administered an opioid medication, subsequently suffer the harm of an opioid-related adverse event, and are administered an opioid antagonist (naloxone) within 12 hours. This measure excludes opioid antagonist (naloxone) administration occurring in the operating room setting.

1b.1. Developer Rationale: Opioids are often the foundation for sedation and pain relief. However, use of opioids can also lead to serious adverse events, including constipation, over sedation, delirium, and respiratory depression. Opioid-related adverse events have both patient-level and financial implications. Patients who experience this event have been noted to have 55% longer lengths of stay, 47% higher costs, 36% higher risk of 30-day readmission, and 3.4 times higher payments than patients without these adverse events (Kessler et al., 2013).

Most opioid-related adverse events are preventable. Of the adverse drug events reported to the Joint Commission´s Sentinel Event database, 47% were due to a wrong medication dose, 29% to improper monitoring, and 11% to other causes (e.g., medication interactions, drug reactions) (Joint Commission, 2012; Overdyk, 2009). Additionally, in a closed-claims analysis, 97% of adverse events were judged preventable with better monitoring and response (Lee et al., 2015) . Naloxone administration is often used as an indicator of a severe opioid-related adverse event, and implementation of this measure can advance safe use of opioids in hospitals and prevent these serious and potentially lethal adverse drug events.

Naloxone is an opioid reversal agent typically used for severe opioid-related adverse events. Naloxone administration has been used in a number of studies as an indicator of opioid-related adverse events (Nwulu et al., 2013; Eckstrand et al., 2009).

From Part 10 of the 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care (Lavonas et al., 2015), the following recommendation is listed for use of Naloxone:

Naloxone is a potent opioid receptor antagonist in the brain, spinal cord, and gastrointestinal system. Naloxone has an excellent safety profile and can rapidly reverse central nervous system (CNS) and respiratory depression in a patient with an opioid-associated resuscitative emergency.

References:

Eckstrand, J. A., Habib, A. S., Williamson, A., Horvath, M. M., Gattis, K. G., Cozart, H., & Ferranti, J. Computerized surveillance of opioid-related adverse drug events in perioperative care: a cross-sectional study. Patient Saf Surg. 2009;3(1), 18.

Kessler ER, Shah M, Gruschkus SK, Raju A. Cost and quality implications of opioid-based postsurgical pain control using administrative claims data from a large health system: opioid-related adverse events and their impact on clinical and economic outcomes. Pharmacotherapy. 2013;33(4):383-391.

Lavonas EJ, Drennan IR, Gabrielli A, Heffner AC, Hoyte CO, Orkin AM, Sawyer KN, Donnino MW. Part 10: Special Circumstances of Resuscitation: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2015 Nov 3;132(18 Suppl 2):S501-18. doi: 10.1161/CIR.0000000000000264. Erratum in: Circulation. 2016 Aug 30;134(9):e122.

Lee, L. A., Caplan, R. A., Stephens, L. S., Posner, K. L., Terman, G. W., Voepel-Lewis, T., & Domino, K. B. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology. 2015:122(3), 659-665.

Nwulu, U., Nirantharakumar, K., Odesanya, R., McDowell, S. E., & Coleman, J. J. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. Eur J Clin Pharmacol. 2013;69(2), 255-259.

Overdyk FJ: Postoperative respiratory depression and opioids. Initiatives in Safe Patient Care, Saxe Healthcare Communications, 2009

The Joint Commission. Safe use of opioids in hospitals. Sentinel Event Alert. 2012(49):1-5. https://www.jointcommission.org/-/media/deprecated-unorganized/imported-assets/tjc/system-folders/topics-library/sea_49_opioids_8_2_12_finalpdf.pdf?db=web&hash=0135F306FCB10D919CF7572ECCC65C84

S.4. Numerator Statement: Inpatient hospitalizations where an opioid antagonist (naloxone) was administered outside of the operating room and within 12 hours following administration of an opioid medication. Only one numerator event is counted per encounter.

S.6. Denominator Statement: Inpatient hospitalizations for patients 18 years or older during which at least one opioid medication was administered. An inpatient hospitalization includes time spent in the emergency department or in observation status when the patients are ultimately admitted to inpatient status.

S.8. Denominator Exclusions: N/A; there are no denominator exclusions

De.1. Measure Type: Outcome

S.17. Data Source: Electronic Health Records

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall, less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

ORAE_Evidence_Form_Final.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

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

Measure Title: Hospital Harm - Opioid-Related Adverse Events

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: **4/2/2021**

1a.1. This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Outcome: **Opioid-Related Adverse Event**

☐ Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

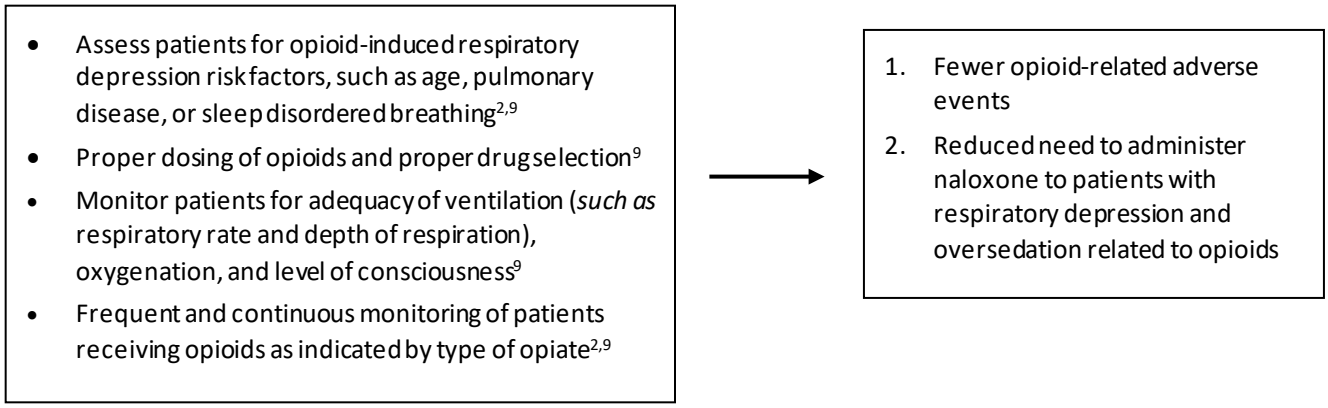
☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The goal of the Opioid-Related Adverse Events electronic clinical quality measure (eCQM) is to improve safety for patients who receive opioids during their hospitalization. Patients who receive excessive doses of opioids (overdose) in the hospital experience confusion, altered consciousness, delirium, respiratory depression, anoxia, anoxic organ damage, and even death as a result.[1-4] The reversal of opioid overdose using an antagonist like naloxone, while necessary to avoid severe effects, in itself causes patients to experience symptoms such as sudden and severe return of pain, nausea, vomiting, tachycardia, seizures, and even cardiac arrest.[5] For these reasons, opioid-related adverse events resulting from opioids administered in the hospital by clinicians are outcomes that should be avoided. Most opioid-related adverse events are preventable with appropriate dosing, patient monitoring, and early response which allows clinicians to reduce opioid dosage before antagonist reversal is necessary.[6]

This eCQM captures the proportion of hospitalized patients aged 18 years and older who suffer the harm of an opioid-related adverse event, defined as receiving a narcotic antagonist (naloxone), with evidence of hospital administration of opioids prior to administration of naloxone. Naloxone administration has been used in studies of computerized adverse drug event surveillance as an indicator of severe opioid-related adverse events.[7,8] By encouraging hospitals to implement evidence-based practices such as routine patient monitoring for potential adverse effects of opioids, this eCQM can lead to better quality of care associated with excessive opioid administration in the hospital setting.[9,10]



References:

1. Kessler ER, Shah M, Gruschkus SK, Raju A. Cost and quality implications of opioid-based postsurgical pain control using administrative claims data from a large health system: opioid-related adverse events and their impact on clinical and economic outcomes. *Pharmacotherapy.* 2013;33(4):383-391.
2. Jungquist CR, Quinlan-Colwell A, Vallerand A, et al. American Society for Pain Management Nursing Guidelines on Monitoring for Opioid-Induced Advancing Sedation and Respiratory Depression: Revisions. Pain Manag Nurs. 2020 Feb;21(1):7-25. Epub 2019 Jul 31.
3. Ramachandran SK, Haider N, Saran KA, et al. Life-threatening critical respiratory events: a retrospective study of postoperative patients found unresponsive during analgesic therapy. *Journal of Clinical Anesthesia.* 2011;23(3):207-213.
4. Dahan A, Aarts L, Smith TW. Incidence, Reversal, and Prevention of Opioid-induced Respiratory Depression. *Anesthesiology*. 2010;112(1):226-238.
5. Wermeling DP. Review of naloxone safety for opioid overdose: practical considerations for new technology and expanded public access. *Ther Adv Drug Saf.* 2015;6(1):20-31.
6. Lee, L. A., Caplan, R. A., Stephens, L. S., Posner, K. L., Terman, G. W., Voepel-Lewis, T., & Domino, K. B. Postoperative opioid-induced respiratory depression: a closed claims analysis. *Anesthesiology*. 2015;122(3), 659-665.
7. Nwulu, U., Nirantharakumar, K., Odesanya, R., McDowell, S. E., & Coleman, J. J. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. *Eur J Clin Pharmacol*. 2013;69(2), 255-259.
8. Eckstrand, J. A., Habib, A. S., Williamson, A., Horvath, M. M., Gattis, K. G., Cozart, H., & Ferranti, J. Computerized surveillance of opioid-related adverse drug events in perioperative care: a cross-sectional study. *Patient Saf Surg*. 2009;3(1), 18.

9. Practice Guidelines for the Prevention, Detection, and Management of Respiratory Depression Associated with Neuraxial Opioid Administration: An Updated Report by the American Society of Anesthesiologists Task Force on Neuraxial Opioids and the American Society of Regional Anesthesia and Pain Medicine. Anesthesiology. 2016 Mar;124(3):535-52. .
10. Lee LA, Caplan RA, Stephens LS, et al. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology. 2015;122(3):659-665.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Best practices to prevent opioid-related adverse events in hospitals have been a major focus by The Joint Commission (TJC), the Institute for Healthcare Improvement (IHI), and the Centers for Medicaid and Medicare Services (CMS).[1-3] Of the opioid-related adverse drug events reported to the Joint Commission's Sentinel Event database, 47% were due to a wrong medication dose, 29% to improper monitoring, and 11% to other causes (for example, medication interactions and drug reactions).[6] Additionally, in a closed-claims analysis, 97% of adverse events were judged preventable with better monitoring and response.[7] Clinical practice guidelines recommend better patient monitoring to improve the measure outcome and reduce the number of opioid-related adverse events.

While monitoring is key to the prevention of opioid-related adverse events, there is considerable variability in hospital monitoring practices. A 2013 study surveyed nurses from 90 institutions in the U.S. and found that pulse oximetry monitoring was more common than other monitoring methods for opioid-induced sedation and respiratory depression.[8] Nonetheless, only about 58% reported using intermittent pulse oximetry and only 25% used continuous monitoring for patient controlled analgesia (PCA). End-tidal carbon dioxide (ETCO2) monitoring was only used for 2.2% of patients on epidural therapy and 1.5% for PCA patients.[8] One hospital found a five-fold reduction in opioid-induced over sedation and respiratory depression cases after implementing targeted interventions such as enhanced monitoring for sedation, improved clinical decision support in the electronic medical record (EMR), and various adjustments to dosing for high-risk patients that included clinician education.[9] Thus, there is room for improvement in monitoring hospitalized patients taking opioids to avoid unintended over sedation and possible opioid-related adverse events.

One study evaluated monitoring practices for patients receiving intravenous (IV) opioids via PCA and found that none of the patients monitored frequently (at least every 2.5 hours) received naloxone in the hospital.[10] Thus, better patient monitoring and response is linked to reduced naloxone administration, signaling avoidance of opioid-related adverse events in the hospital.

References:

1. The Joint Commission. Joint Commission enhances pain assessment and management requirements for accredited hospitals. 2017;37 (7) 1-4.

https://www.jointcommission.org/-/media/tjc/documents/standards/jc-requirements/2018-requirements/joint_commission_enhances_pain_assessment_and_management_requirements_for_accredited_hospitals1.pdf?db=web&hash=1DFAA78F3C6EDD8AA2A1A152D18D4409

2. Institute for Healthcare Improvement. (2012). How to guide: Prevent harm from high alert medications. Cambridge, MA. http://www.ihi.org/resources/Pages/Tools/HowtoGuidePreventHarmfromHighAlertMedications.aspx
3. Centers for Medicare and Medicaid Services. (2014). Requirements for hospital medication administration, particularly intravenous (IV) medications and postoperative care of patients receiving IV opioids. Baltimore, MD. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/SurveyCertificationGenInfo/Downloads/Survey-and-Cert-Letter-14-15.pdf
4. Jungquist CR, Quinlan-Colwell A, Vallerand A, et al. American Society for Pain Management Nursing Guidelines on Monitoring for Opioid-Induced Advancing Sedation and Respiratory Depression: Revisions. Pain Manag Nurs. 2020 Feb;21(1):7-25. Epub 2019 Jul 31.
5. Practice Guidelines for the Prevention, Detection, and Management of Respiratory Depression Associated with Neuraxial Opioid Administration: An Updated Report by the American Society of Anesthesiologists Task Force on Neuraxial Opioids and the American Society of Regional Anesthesia and Pain Medicine. Anesthesiology. 2016 Mar;124(3):535-52
6. The Joint Commission. Safe use of opioids in hospitals. Sentinel Event Alert. 2012(49):1-5.
7. Lee LA, Caplan RA, Stephens LS, et al. Postoperative opioid-induced respiratory depression: a closed claims analysis. *Anesthesiology*. 2015;122(3):659-665.
8. Willens JS, Jungquist CR, Cohen A, Polomano R. ASPMN survey--nurses' practice patterns related to monitoring and preventing respiratory depression. *Pain Management Nursing.* 2013;14(1):60-65.
9. Meisenberg B, Ness J, Rao S, Rhule J, Ley C. Implementation of solutions to reduce opioid-induced oversedation and respiratory depression. Am J Health Syst Pharm. 2017;74:162-169.
10. Jungquist CR, Correll DJ, Fleisher LA, et al. Avoiding Adverse Events Secondary to Opioid-Induced Respiratory Depression: Implications for Nurse Executives and Patient Safety. *Journal of Nursing Administration.* 2016;46(2):87-94.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE *(e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.*

Opioids are often the foundation for sedation and pain relief. However, use of opioids can also lead to serious adverse events, including constipation, over sedation, delirium, and respiratory depression.

Opioid-related adverse events have both patient-level and financial implications. Patients who experience this event have been noted to have 55% longer lengths of stay, 47% higher costs, 36% higher risk of 30-day readmission, and 3.4 times higher payments than patients without these adverse events (Kessler et al., 2013).

Most opioid-related adverse events are preventable. Of the adverse drug events reported to the Joint Commission´s Sentinel Event database, 47% were due to a wrong medication dose, 29% to improper monitoring, and 11% to other causes (e.g., medication interactions, drug reactions) (Joint Commission, 2012; Overdyk, 2009). Additionally, in a closed-claims analysis, 97% of adverse events were judged preventable with better monitoring and response (Lee et al., 2015) . Naloxone administration is often used as an indicator of a severe opioid-related adverse event, and implementation of this measure can advance safe use of opioids in hospitals and prevent these serious and potentially lethal adverse drug events.

Naloxone is an opioid reversal agent typically used for severe opioid-related adverse events. Naloxone administration has been used in a number of studies as an indicator of opioid-related adverse events (Nwulu et al., 2013; Eckstrand et al., 2009).

From Part 10 of the 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care (Lavonas et al., 2015), the following recommendation is listed for use of Naloxone:

Naloxone is a potent opioid receptor antagonist in the brain, spinal cord, and gastrointestinal system. Naloxone has an excellent safety profile and can rapidly reverse central nervous system (CNS) and respiratory depression in a patient with an opioid-associated resuscitative emergency.

References:

Eckstrand, J. A., Habib, A. S., Williamson, A., Horvath, M. M., Gattis, K. G., Cozart, H., & Ferranti, J. Computerized surveillance of opioid-related adverse drug events in perioperative care: a cross-sectional study. Patient Saf Surg. 2009;3(1), 18.

Kessler ER, Shah M, Gruschkus SK, Raju A. Cost and quality implications of opioid-based postsurgical pain control using administrative claims data from a large health system: opioid-related adverse events and their impact on clinical and economic outcomes. Pharmacotherapy. 2013;33(4):383-391.

Lavonas EJ, Drennan IR, Gabrielli A, Heffner AC, Hoyte CO, Orkin AM, Sawyer KN, Donnino MW. Part 10: Special Circumstances of Resuscitation: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2015 Nov 3;132(18 Suppl 2):S501-18. doi: 10.1161/CIR.0000000000000264. Erratum in: Circulation. 2016 Aug 30;134(9):e122.

Lee, L. A., Caplan, R. A., Stephens, L. S., Posner, K. L., Terman, G. W., Voepel-Lewis, T., & Domino, K. B. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology. 2015:122(3), 659-665.

Nwulu, U., Nirantharakumar, K., Odesanya, R., McDowell, S. E., & Coleman, J. J. Improvement in the detection of adverse drug events by the use of electronic health and prescription records: an evaluation of two trigger tools. Eur J Clin Pharmacol. 2013;69(2), 255-259.

Overdyk FJ: Postoperative respiratory depression and opioids. Initiatives in Safe Patient Care, Saxe Healthcare Communications, 2009

The Joint Commission. Safe use of opioids in hospitals. Sentinel Event Alert. 2012(49):1-5. https://www.jointcommission.org/-/media/deprecated-unorganized/imported-assets/tjc/system-folders/topics-library/sea_49_opioids_8_2_12_finalpdf.pdf?db=web&hash=0135F306FCB10D919CF7572ECCC65C84

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. ***(This is required for maintenance of endorsement****. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

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. A detailed breakdown of the characteristics of the measured facilities and the patient population can be found in the attached Measure Testing Form section 1.7 (Beta Dataset – Implementation Testing).

The measure performance, including the denominator, numerator, and measure rate by hospital, are as follows:

Hospital Test Site 1 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019 - 12/31/2019
- Denominator(encounters): 1,839:
- Numerator: 2
- Performance rate: 0.11%
- Standard deviation: 3.30%
- 95% confidence interval: [0%, 0.26%]

Hospital Test Site 2 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019 - 12/31/2019
- Denominator (encounters): 2,089
- Numerator: 7
- Performance rate: 0.34%
- Standard deviation: 5.78%
- 95% confidence interval: [0.09%, 0.58%]

Hospital Test Site 3 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019-12/31/2019
- Denominator: 1,784
- Numerator: 8
- Performance Rate: 0.45%
- Standard deviation: 6.68%
- 95% confidence interval: [0.14%, 0.76%]

Hospital Test Site 4 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019-12/31/2019
- Denominator: 11,273
- Numerator: 50
- Performance Rate: 0.45%
- Standard deviation: 6.71%
- 95% confidence interval: [0.33%, 0.58%]

Hospital Test Site 5 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019-12/31/2019
- Denominator: 13,307
- Numerator: 44
- Performance Rate: 0.33%
- Standard deviation: 5.74%
- 95% confidence interval: [0.23%, 0.43%]

Hospital Test Site 6 (Beta Dataset – Implementation Testing per Testing Form)

- Data collection period: 1/1/2019-12/31/2019
- Denominator: 18,425
- Numerator: 64
- Performance Rate: 0.35%
- Standard deviation: 5.88%
- 95% confidence interval: [0.26%, 0.43%]

Overall Performance (calculated at the hospital level)

- Number of hospitals: 6
- Performance rate: 0.34%
- Standard deviation: 0.12%
- Range: 0.11%-0.45%

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

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.

ORAE 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. Also, larger hospitals (e.g., test sites 4 to 6), though having more numerator admissions, do not necessarily have higher ORAE rates. This

suggests that all hospitals, irrespective of size, need to follow best practices in patient care to prevent ORAE.

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.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(**This is required for maintenance of endorsement**. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

We examined the measure performance rate in various subgroups of population. Of note, these summary statistics are calculated at the encounter-level and derived from a sample of six hospitals and may not be generalizable to the entire population. Data below are from initial development testing; this eCQM is not yet implemented. The measure performance was stratified for disparities by age, sex, race, ethnicity, and payer source for each of the six Beta Implementation Test Sites as presented in Tables 15-16 in the testing attachment. In addition, summary statistics that including the mean performance rate and standard deviation for each demographic characteristic across all six Beta Implementation Test Sites are provided below.

Demographic Characteristic//Mean Rate//Std. Dev//95%CI

Across all denominator patient-encounters//0.36%//6.00%//[0.31%, 0.41%]

Sub-groups

Age bins

18-35//0.07%//2.60%//[0.02%, 0.11%]

36-64//0.41%//6.40%//[0.32%, 0.50%]

65+//0.51%//7.11%//[0.40%, 0.62%]

Sex

Male//0.42%//5.63%//[0.33%, 0.51%]

Female//0.32%//6.50%//[0.25%, 0.38%]

Race

Black or African American//0.24%//4.94%//[0.08%, 0.40%]

White//0.37%//6.04%//[0.30%, 0.43%]

Other//0.48%//6.91%//[0.29%, 0.67%]

Unknown//0.25%//4.97%//[0.08%, 0.42%]

Ethnicity

Hispanic or Latino//0.32%//5.65%//[0.21%, 0.43%]

Non-Hispanic//0.40%//6.28%//[0.33%, 0.46%]

Unknown//0.25%//5.01%//[0.12%, 0.38%]

(Primary) Payer

Medicare//0.52%//7.21%//[0.41%, 0.63%]

Medicaid//0.29%//5.36%//[0.21%, 0.37%]

Private Insurance//0.26%//5.09%//[0.15%, 0.37%]

Self-pay or Uninsured//0.24%//4.85%//[0.07%, 0.40%]

Other//0.40%//6.35%//[0.08%, 0.73%]

Unknown//0.00%//0.00%//N/A

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, **as specified**, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

Final measure specifications for implementation will be made publicly available on CMS' appropriate quality reporting website, once finalized through the NQF endorsement and CMS rulemaking processes.

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is an eMeasure **Attachment:** Opioid_v6_02_Artifacts.zip,ORAE-_Bonnie_v4.2.0__Measure_View_-_CMS819v0.pdf

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **:** Opioid_Value_Set_Directory.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

N/A

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

***IF an OUTCOME MEASURE**, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

Inpatient hospitalizations where an opioid antagonist (naloxone) was administered outside of the operating room and within 12 hours following administration of an opioid medication. Only one numerator event is counted per encounter.

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

***IF an OUTCOME MEASURE,** describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

This is an eCQM, and therefore uses electronic health record data to calculate the measure score. The time period for data collection is during an inpatient hospitalization, beginning at hospital arrival (whether through emergency department, observation stay, or directly admitted as inpatient).

All data elements necessary to calculate this measure are defined within value sets available in the Value Set Authority Center (VSAC), and listed below.

The Opioid antagonist (naloxone) is defined by the value set Opioid Antagonist (2.16.840.1.113752.1.4.1179.1).

Opioids are defined by the value set Opioids, All (2.16.840.1.113762.1.4.1196.226).

The location for opioid administration is defined by the code Operating Room/Suite (HSLOC Code 1096-7).

To access the value sets for the measure, please visit the Value Set Authority Center (VSAC), sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Inpatient hospitalizations for patients 18 years or older during which at least one opioid medication was administered. An inpatient hospitalization includes time spent in the emergency department or in observation status when the patients are ultimately admitted to inpatient status.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection*

items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)

***IF an OUTCOME MEASURE**, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

This measure includes all patients aged 18 years and older at the time of admission, and all payers. Measurement period is one year. This measure is at the hospital admission level; only one numerator event is counted per encounter.

Inpatient Encounters are represented using the value set of Encounter Inpatient (2.16.840.1.113883.3.666.5.307).

Emergency Department visits are represented using the value set of Emergency Department Visit (2.16.840.1.113883.3.117.1.7.1.292).

Patients who had observation encounters are represented using the value set of Observation Services (2.16.840.1.113762.1.4.1111.143).

Opioids are defined by the value set Opioids, All (2.16.840.1.113762.1.4.1196.226).

To access the value sets for the measure, please visit the Value Set Authority Center, sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

N/A; there are no denominator exclusions

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

N/A

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

N/A; this measure is not stratified.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

This measure defines the indication of a harm for an opioid-related adverse event by assessing administration of an opioid antagonist (naloxone).

To calculate the hospital-level measure result, divide the total numerator events by the total number of qualifying encounters (denominator).

Qualifying encounters (denominator) include all patients 18 years of age or older at the start of the encounter with at least one opioid medication administered during the encounter.

To create the numerator:

1. First, start with those encounters meeting denominator criteria

2. Next, remove all events where an opioid antagonist (naloxone) was only administered in the operating room.

Opioid antagonist administrations in the operating room are excluded because they could be part be part of the sedation plan as administered by an anesthesiologist. Encounters that include use of opioid antagonists for procedures and recovery outside of the operating room (e.g., bone marrow biopsy and PACU) are included in the numerator, as it would indicate the patient was over-sedated. Note that should a facility not utilize temporary patient locations, alternative times may be used to determine whether a patient is in the operating room during opioid antagonist administration. Since anesthesia end time could represent the time that the anesthesiologist signed off, and thus may include the patient´s time in the PACU, this should be avoided.

3. Finally, remove all administrations of naloxone that were given greater than 12 hours following hospital administration of an opioid medication .

This eCQM is an episode-based measure.

This version of the eCQM uses QDM version 5.5. Please refer to the eCQI resource center (https://ecqi.healthit.gov/qdm) for more information on the QDM.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A; this measure does not use a sample.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A; this measure does not use a survey.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Electronic Health Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Hospitals collect EHR data using certified electronic health record technology (CEHRT). The MAT output, which includes the human readable and XML artifacts of the clinical quality language (CQL) for the measure are contained in the eCQM specifications attached. No additional tools are used for data collection for eCQMs.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

N/A

2. Validity – See attached Measure Testing Submission Form

ORAE_NQF_Testing_Attachment_v2.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1, 2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST

use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

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:

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

*cell intentionally left blank

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:	☐ other:

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:	☐ other:

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

Dataset	Applicable Section in the Testing Attachment	Description of Dataset	EHR Vendor
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

Dataset	Applicable Section in the Testing Attachment	Description of Dataset	EHR Vendor
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 1. Demographic Characteristics of Measure Initial Population (Site 1-3)

Initial Population Characteristics	Test Site 1: n	Test Site 1: %	Test Site 2: n	Test Site 2: %	Test Site 3: n	Test Site 3: %
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%
Age bins: 36-64	500	33%	629	33%	473	31%
Age bins: 65+	607	39%	845	45%	799	52%
Sex: Male	513	33%	612	32%	568	37%
Sex: Female	1,024	67%	1,277	68%	976	63%
Race: Black or African-American	39	3%	80	4%	15	1%
Race: 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%
Ethnicity: Non-Hispanic	0	0%	1,810	96%	1,534	99%

Initial Population Characteristics	Test Site 1: n	Test Site 1: %	Test Site 2: n	Test Site 2: %	Test Site 3: n	Test Site 3: %
Ethnicity: Unknown	1,537	100%	5	0.3%	0	0%
(Primary) Payer: Medicare	718	47%	1,025	54%	849	55%
(Primary) Payer: Medicaid	395	26%	377	20%	191	12%
(Primary) Payer: Private Insurance	292	19%	314	17%	456	30%
(Primary) Payer: Self-pay or Uninsured	1	0%	19	1%	12	1%
(Primary) Payer: Other†	116	8%	62	3%	20	1%
(Primary) Payer: 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.

*cell intentionally left blank

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

Initial Population Characteristics	Test Site 4: n	Test Site 4: %	Test Site 5: n	Test Site 5: %	Test Site 6: n	Test Site 6: %
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%
Age bins: 36-64	3,418	36%	4,415	41%	6,316	41%
Age bins: 65+	2,861	30%	4,037	37%	4,653	30%

Initial Population Characteristics	Test Site 4: n	Test Site 4: %	Test Site 5: n	Test Site 5: %	Test Site 6: n	Test Site 6: %
Sex: Male	3,226	34%	4,412	41%	6,473	42%
Sex: Female	6,187	66%	6,415	59%	8,788	58%
Race: Black or African American	564	6%	780	7%	1,368	9%
Race: White	7,034	75%	6,708	62%	12,002	79%
Race: Other	1,514	16%	1,459	13%	1,304	9%
Race: Unknown	301	3%	1,880	17%	587	4%
Ethnicity: Hispanic or Latino	4,033	43%	984	9%	4,182	27%
Ethnicity: Non-Hispanic	5,140	55%	7,832	72%	9,853	65%
Ethnicity: Unknown	240	3%	2,011	18.6%	1,226	8%
(Primary) Payer: Medicare	2,174	23%	2,900	27%	5,399	35%
(Primary) Payer: Medicaid	4,729	50%	4,471	41%	5,121	34%
(Primary) Payer: Private Insurance	1,206	13%	1,859	17%	3,600	24%
(Primary) Payer: Self-pay or Uninsured	1,093	12%	1,348	12%	524	3%
(Primary) Payer: Other	210	2%	249	2%	617	4%
(Primary) Payer: 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.

*cell intentionally left blank

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: 1	Rater B: 2	Total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

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:

$$p_e = \frac{n_{1.}}{n} \times \frac{n_{.1}}{n} + \frac{n_{2.}}{n} \times \frac{n_{.2}}{n}$$

then Cohen's Kappa coefficient is equal to $\kappa = \frac{p_0 - p_e}{1 - p_e}$, where $p_0 = \frac{n_{11} + n_{22}}{n}$ 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

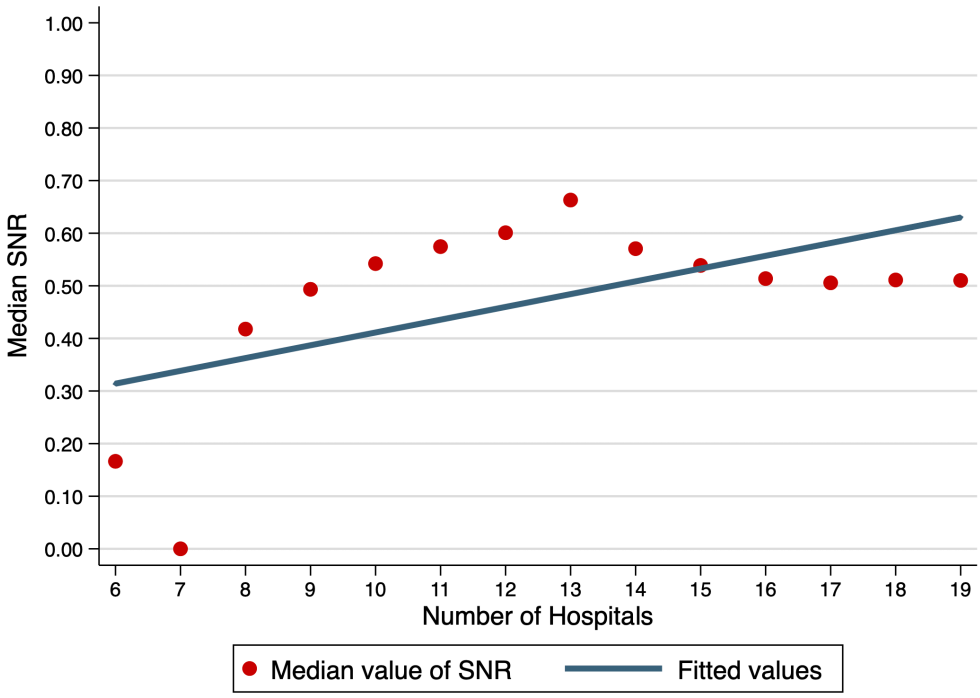
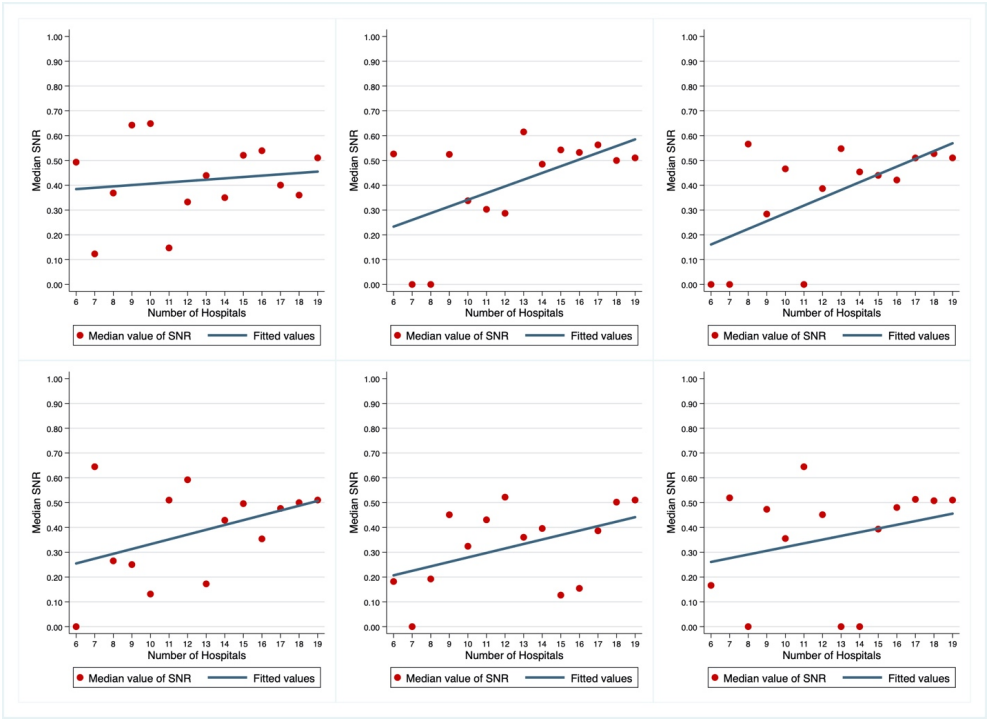


Figure 2. Relationship Between the Number of Hospitals and the Median Value of SNR



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)

	Test Site 1			Test Site 2			Test Site 3		
Data Element	**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)

	Test Site 4			Test Site 5			Test Site 6		
Data Element	**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: Kappa	Test Site 2: Kappa	Test Site 3: Kappa	Test Site 4: Kappa	Test Site 5: Kappa	Test Site 6: Kappa
Patients 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:

$$n = \frac{z_{\frac{\alpha}{2}}^2 \cdot p \cdot (1-p)}{moe^2}$$

, where a denotes the type I error rate, moe denotes the margin of error, and p is PPV. We simulated a series of moe and target p values for MRSS and 95% confidence interval (CI). For example, with a moe 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

Data Element	Test Site 1: Count	Test Site 2: Count	Test Site 3: Count	Test Site 4: Count	Test Site 5: Count	Test Site 6: 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)

	Test Site 1			Test Site 2		
Data Element	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**
Patients 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)

	Test Site 3			Test Site 4		
Data Element	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**
Patients 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	8	8	100%	51	51	100%

Data Element	# Cases Matched in EHR (n)	# Cases in Abstraction (n)	Percent Match (%)	# Cases Matched in EHR (n)	# Cases in Abstraction (n)	Percent Match (%)
within 12hrs of the opioid administration and outside of the operating room						

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

	Test Site 5			Test Site 6		
Data Element	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**	**# Cases Matched in EHR (n)**	**# Cases in Abstraction (n)**	**Percent Match (%)**
Patients 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%

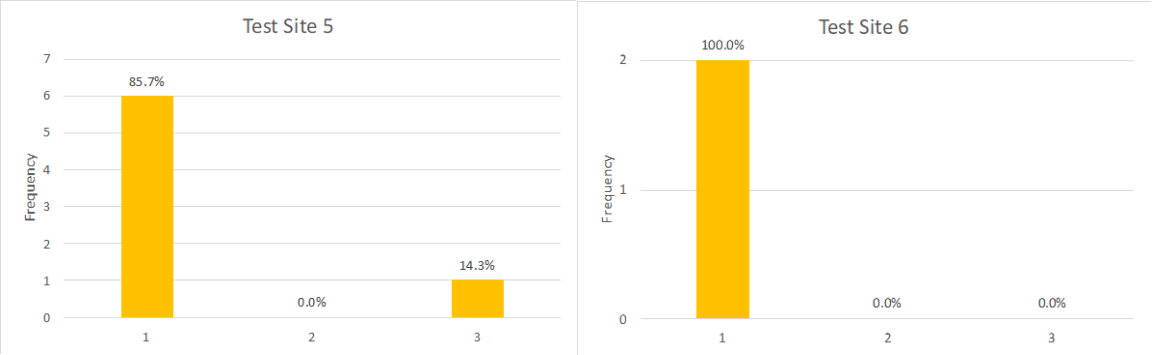
Across the six implementation test sites, all but two data elements showed an agreement rate of 100%, indicating that valid and accurate data were extracted from patients' EHRs. The misaligned case in test site 5 was due to the clinical abstractor review determining the encounter to be a numerator event, 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.

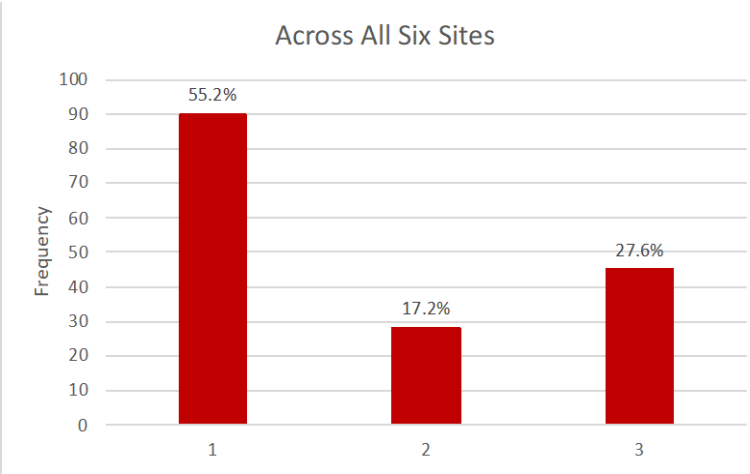
To examine if the numerator cases identified by the quality reporting engine are true positives, clinical abstractors pulled additional information regarding the indication for and subsequent reaction to the naloxone administration from the nurse notes and physician orders. We grouped patient responses 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



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



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)

	Test Site 1			Test Site 2			Test Site 3		
Measure Component	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)

	Test Site 4			Test Site 5			Test Site 6		
Measure Component	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)

	Test Site 1 (N=100)			Test Site 2 (N=100)			Test Site 3 (N=1000		
Measure	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)

	Test Site 4 (N=100)			Test Site 5 (N=100)			Test Site 6 (N=1000		
Measure	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*

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.

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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		
Measure	**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%
Age bins (Sub-groups): 18-35	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.35%	5.90%	0.00%
Age bins (Sub-groups): 36-64	0.16%	3.98%	0.00%	0.42%	6.45%	0.00%	0.73%	8.50%	0.00%
Age bins (Sub-groups): 65+	0.13%	3.61%	0.00%	0.43%	6.52%	0.00%	0.32%	5.63%	0.00%
Sex (Sub-groups): Male	0.30%	5.51%	0.00%	0.59%	7.64%	0.00%	0.15%	3.88%	0.00%
Sex (Sub-groups): Female	0.00%	0.00%	0.00%	0.21%	4.61%	0.00%	0.63%	7.88%	0.00%
Race (Sub-groups): Black or African American	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Race (Sub-groups): White	0.11%	3.35%	0.00%	0.35%	5.93%	0.00%	0.47%	6.81%	0.00%
Race (Sub-groups): Other	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Race (Sub-groups): Unknown	*	*	*	0.00%	*	0.00%	0.00%	0.00%	0.00%
Ethnicity (Sub-groups): Hispanic or Latino	*	*	*	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Ethnicity (Sub-groups): Non-Hispanic	*	*	*	0.35%	5.90%	0.00%	0.45%	6.70%	0.00%
Ethnicity (Sub-groups): Unknown	0.11%	3.30%	0.00%	0.00%	0.00%	0.00%	*	*	*
(Primary) Payer (Sub-groups): Medicare	0.11%	3.30%	0.00%	0.34%	5.86%	0.00%	0.59%	7.65%	0.00%

Measure	Mean	Std.Dev	P50	Mean	Std.Dev	P50	Mean	Std.Dev	P50
(Primary) Payer (Sub-groups): Medicaid	0.22%	4.64%	0.00%	0.00%	0.00%	0.00%	0.45%	6.74%	0.00%
(Primary) Payer (Sub-groups): Private Insurance	0.00%	0.00%	0.00%	0.30%	5.45%	0.00%	0.20%	4.51%	0.00%
(Primary) Payer (Sub-groups): Self-pay or Uninsured	0.00%	*	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
(Primary) Payer (Sub-groups): Other	0.00%	0.00%	0.00%	3.03%	17.27%	0.00%	0.00%	0.00%	0.00%
(Primary) Payer (Sub-groups): Unknown	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

*cell intentionally left blank

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

	Test Site 4			Test Site 5			Test Site 6		
Measure	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%
Age bins (Sub-groups): 18-35	0.03%	1.73%	0.00%	0.08%	2.77%	0.00%	0.08%	2.90%	0.00%
Age bins (Sub-groups): 36-64	0.50%	7.06%	0.00%	0.35%	5.93%	0.00%	0.40%	6.31%	0.00%
Age bins (Sub-groups): 65+	0.79%	8.87%	0.00%	0.44%	6.60%	0.00%	0.49%	7.00%	0.00%
Sex (Sub-groups): Male	0.78%	8.79%	0.00%	0.36%	5.96%	0.00%	0.31%	5.55%	0.00%
Sex (Sub-groups): Female	0.27%	5.14%	0.00%	0.31%	5.58%	0.00%	0.38%	6.13%	0.00%
Race (Sub-groups): Black or African American	0.13%	3.60%	0.00%	0.29%	5.34%	0.00%	0.29%	5.40%	0.00%
Race (Sub-groups): White	0.45%	6.66%	0.00%	0.34%	5.82%	0.00%	0.36%	5.96%	0.00%
Race (Sub-groups): Other	0.54%	7.31%	0.00%	0.51%	7.15%	0.00%	0.39%	6.25%	0.00%
Race (Sub-groups): Unknown	0.87%	9.30%	0.00%	0.18%	4.19%	0.00%	0.16%	4.05%	0.00%

Measure	Mean	Std.Dev	P50	Mean	Std.Dev	P50	Mean	Std.Dev	P50
Ethnicity (Sub-groups): Hispanic or Latino	0.45%	6.67%	0.00%	0.25%	5.01%	0.00%	0.22%	4.72%	0.00%
Ethnicity (Sub-groups): Non-Hispanic	0.45%	6.66%	0.00%	0.35%	5.92%	0.00%	0.40%	6.35%	0.00%
Ethnicity (Sub-groups): Unknown	0.70%	8.38%	0.00%	0.29%	5.36%	0.00%	0.29%	5.35%	0.00%
(Primary) Payer (Sub-groups): Medicare	0.79%	8.83%	0.00%	0.43%	6.52%	0.00%	0.54%	7.36%	0.00%
(Primary) Payer (Sub-groups): Medicaid	0.34%	5.80%	0.00%	0.36%	5.97%	0.00%	0.20%	4.43%	0.00%
(Primary) Payer (Sub-groups): Private Insurance	0.22%	4.68%	0.00%	0.29%	5.34%	0.00%	0.28%	5.31%	0.00%
(Primary) Payer (Sub-groups): Self-pay or Uninsured	0.49%	6.96%	0.00%	0.13%	3.58%	0.00%	0.00%	0.00%	0.00%
(Primary) Payer (Sub-groups): Other	0.42%	6.45%	0.00%	0.00%	0.00%	0.00%	0.41%	6.37%	0.00%
(Primary) Payer (Sub-groups): Unknown	0.00%	0.00%	0.00%	*	*	*	*	*	*

*cell intentionally left blank

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 <u>2b3.9</u>

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 non-responders) 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)

	Test Site 1			Test Site 2			Test Site 3		
Data Element	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)

	Test Site 4			Test Site 5			Test Site 6		
Data Element	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 non-responders) 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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

This is an eCQM that uses all data elements from defined fields in the EHR. Of all sites used for the measure feasibility assessment, some reported that their anesthesiologists document their activities on paper-based anesthesia records inside of the OR rather than via the electronic medication administration record (eMAR). This suggests that, at this time, for these sites, opioid and naloxone administration inside of the OR will not be available for structured electronic extraction or appear in patient EHRs. For opioid and naloxone administration outside of OR suite, however, all test sites confirmed that they are documented in the eMARs, and available for electronic extraction. Test sites' decisions to document opioid administration inside of the OR on paper can be driven by many factors, one of which is a workflow preference. Since all hospital-based EHR vendor systems offer anesthesia modules, there should be no technical limitation in transitioning paper-based documentation to electronic documentation. Given that non-anesthesia-related opioid administrations are already captured electronically, we are optimistic that measure implementation is still feasible. Moreover, measure implementation will drive workflow changes toward electronic capture within the OR.

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment: ORAE_NQF_feasibility_scorecard_vFinal_External.xlsx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

This measure is not instrument-based. Feasibility assessment of this eCQM across twenty-three hospitals with four different EHR vendors (Epic, Cerner, Meditech, and Allscripts) found that the critical data elements used for measure calculation were, for the most part, reliably available in a structured format within the EHR, captured as part of the course of care, accurately recorded information, and coded using nationally accepted terminology.

However, some sites reported that their anesthesiologists document activities on paper-based anesthesia records inside of the OR rather than via the electronic medication administration record (eMAR). Since all hospital-based EHR vendor systems offer anesthesia modules, there should be no technical limitation in transitioning paper-based documentation to electronic documentation. Given that non-anesthesia-related opioid administrations are already captured electronically, we are optimistic that measure implementation is still feasible. Moreover, measure implementation will drive workflow changes toward electronic capture within the OR. Of all the test sites, 21 confirmed that their EHR systems are capable of collecting such information and documenting the events either directly in patient EHRs using encounter location or via proxy information, such as the location associated with nurse administration of medication or time into and out of the OR.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees associated with the use of this eCQM. Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services (CMS). Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html).

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Not in use	*

*cell intentionally left blank

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

N/A; this eCQM is under initial endorsement review and is not currently used in any accountability program.

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This eCQM is under initial endorsement review and is not currently used in any accountability program. In December 2017, this measure was presented to the Measure Applications Partnership (MAP), who recommended this measure be revised and resubmitted prior to rulemaking. The MAP asked the measure developers to demonstrate reliability and validity in their completed testing, and submit the finalized eCQM to NQF for review and endorsement; this had been completed and submitted to NQF in the Spring 2019 cycle. Based on feedback received from NQF during the 2019 Spring cycle, CMS has subsequently made substantive updates and re-tested the measure. CMS intends to submit this eCQM for the 2021-2022 pre-rulemaking process including the Measures Under Consideration list and the Measures Application Partnership (MAP). Following MAP 2021-2022 review, we envision that this measure will be considered for accountability programs via future rulemaking. Thus, CMS is seeking endorsement by NQF.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Following MAP 2021-2022 review, we envision that this measure will be considered for accountability programs via future rulemaking.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

N/A; this measure is being submitted as de novo as has not yet been implemented. Implementation is planned pending finalization of the NQF endorsement and CMS rulemaking processes.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

N/A; this measure is being submitted as de novo as has not yet been implemented. Implementation is planned pending finalization of the NQF endorsement and CMS rulemaking processes.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

N/A; this measure is being submitted as de novo as has not yet been implemented. Implementation is planned pending finalization of the NQF endorsement and CMS rulemaking processes.

4a2.2.2. Summarize the feedback obtained from those being measured.

N/A; this measure is being submitted as de novo as has not yet been implemented. Implementation is planned pending finalization of the NQF endorsement and CMS rulemaking processes.

4a2.2.3. Summarize the feedback obtained from other users

While this measure does not have usability information from measured entities, as it is being developed de novo and has not been implemented yet, our team sought input from multiple stakeholder groups throughout the measure development process. We believe in a transparent measure development process, and highly value the feedback received on the measure. During development, a technical expert panel composed of a variety of stakeholders was engaged at various stages of development to obtain balanced, expert input. We also solicited and received feedback on the measure through an MMS Blueprint 44-day Public Input Period during development.

We also received feedback from various stakeholders in 2019 including the NQF Patient Safety Standing Committee during the Spring 2019 cycle as well as a 60-day comment period for Federal rulemaking. Concerns raised by commenters included how the measure was specified as a proportion of all hospitalized patients, the use of naloxone as an indicator for quality, potential unintended consequences, and a potential lack of a performance gap. All of these concerns have been addressed by the revisions to the measure specifications described in the following section (4a2.3).

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

As noted above, input received from TEP members was instrumental to the development and specification of this measure. Feedback received during the NQF 2019 Spring cycle and public comment during Federal rulemaking was also incorporated into the measure refinement and re-testing process. Specifically:

- We updated the measure value sets to ensure that the most current codes hospital administered opioids and naloxone are used and that the codes harmonize across other eCQMs in current CMS quality reporting programs;

- We limited the measure denominator to encounters where patients received at least one opioid during the hospitalization;

- We added a 12-hour time window such that the opioid administration must precede the subsequent naloxone administration to ensure that a hospital administered opioid was the cause for the naloxone administration;

- We subsequently re-tested the refined measure for feasibility at 23 hospital test sites using four EHR vendors (Epic, Cerner, Meditech; and Allscripts);

- We also re-tested for the scientific acceptability of the measure's properties including reliability and validity at six beta implementation test sites.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

This is a new eCQM and there is no time trend information available regarding facility performance improvement. This eCQM is not currently used in any quality improvement program, but a primary goal of the eCQM is to provide hospitals with performance information necessary to implement focused quality improvement efforts.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

We did not identify any unintended consequences during eCQM development or testing. However, CMS is committed to monitoring this eCQM's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care and other negative unintended consequences for patients. However, it is important that the eCQM, as currently specified, does not detect false positives. To verify this, we conducted empirical tests to examine whether numerator cases identified by the measure are true positives. In the chart review (or parallel-form comparison) process, we instructed clinical abstractors to extract both indications for and patient subsequent responses to the naloxone administration. We found that the predominant rationale for subsequent naloxone administration was that patients were somnolent or unresponsive, with the second mostly cited reason being opiate reversal. In terms of patient responses to naloxone administration, we found that the most frequently documented was: patient showed clear signs of response to naloxone administration. This qualitative evidence solidifies the evaluation of measure logic and suggests that the measure can correctly predict a true positive.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

No unexpected benefits were noted during eCQM development or testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria **and** there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

3316: Safe Use of Opioids – Concurrent Prescribing

3389: Concurrent Use of Opioids and Benzodiazepines (COB)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

The Hospital Harm – Opioid Related Adverse Events eCQM, the Safe Use of Opioids – Concurrent Prescribing Measure (NQF #3316e), and the Concurrent Use of Opioids and Benzodiazepines (NQF #3389) all have the same general target population, which are adult patients who receive opioids. However, the focus of each measure is very different. The Hospital Harm – Opioid Related Adverse Events eCQM focuses on patients who receive excessive doses of opioids during their hospitalization and, subsequently, require naloxone to prevent further patient harm. In contrast, NQF #3316e focuses on patients who receive concurrent opioid or opioid and benzodiazepine prescriptions at discharge, putting them at-risk of adverse drug events after hospital discharge, and NQF #3389 tracks concurrent opioid and benzodiazepine outpatient prescriptions. As a result of the varying measure focuses, the Hospital Harm – Opioid Related Adverse Events eCQM has a broad denominator of all inpatient adults >=18 years who received a hospital administered opioid, while NQF #3316e has a more narrow denominator of adults >=18 years prescribed an opioid or benzodiazepine at discharge from a hospital-based encounter. NQF #3316e also excludes patients with an active cancer diagnoses, palliative care order, or length of stay >120 days. NQF #3389 addresses outpatient prescription claims and excludes patients in hospice, or with a cancer or sickle cell disease diagnosis.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Annese, Abdullah-Mclaughlin, Annese Abdullah-Mclaughlin@cms.hhs.gov, 410-786-2995-

Co.3 Measure Developer if different from Measure Steward: IMPAQ International, LLC

Co.4 Point of Contact: Katie, Magoulick, nqf@impaqint.com, 443-259-5449-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

TEP Members:

David Baker, MD, MPH; The Joint Commission

Cynthia Barnard, PhD, MBA, MSJS; Northwestern Memorial HealthCare

Lisa Freeman, Connecticut Center for Patient Safety

Christine Norton, MA; Consumer/Patient Caregiver

David Hopkins, MS, PhD; Stanford University

Kevin Kavanagh, MD, MS; Health Watch USA

Joseph Kunisch, PhD, RN-BC, CPHQ, Memorial Hermann Hospital System

Timothy Lowe, PhD; Premier, Inc.

Amita Rastogi, MD, MHA, CHE, MS; Remedy Partners

Karen Zimmer, MD, MPH; Jefferson School of Population Health and Jefferson University College of Medicine

Steven Jarrett, Pharm.D., Atrium Health

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure? As a de novo measure submission, we anticipate annual updates and potentially triennial endorsement

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: Limited proprietary coding is contained in the Measure specifications for user convenience. Users of proprietary code sets should obtain all necessary licenses from the owners of the code sets. IMPAQ disclaims all liability for use or accuracy of any third party codes contained in the specifications. CPT(R) contained in the Measure specifications is copyright 2004-2020 American Medical Association. LOINC(R) copyright 2004-2020 Regenstrief Institute, Inc. This material contains SNOMED Clinical Terms(R) (SNOMED CT[R]) copyright 2004-2020 International Health Terminology Standards Development Organisation. ICD-10 copyright 2020 World Health Organization. All Rights Reserved.

Ad.7 Disclaimers: This measure and specifications are subject to further revisions. This performance measure is not a clinical guideline and does not establish a standard of medical care, and has not been tested for all potential applications. THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. Due to technical limitations, registered trademarks are indicated by (R) or [R] and unregistered trademarks are indicated by (TM) or [TM].

Ad.8 Additional Information/Comments: This measure was originally developed, specified, and tested by Yale New Haven Health Service Corporation Center for Outcomes Research and Evaluation, and by Mathematica Policy Research on behalf of the Centers for Medicare and Medicaid Services (CMS). IMPAQ International, LLC assumed developer responsibility for this measure in March 2019.