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

Brief Measure Information

NQF #: 3316
Measure Title: Safe Use of Opioids – Concurrent Prescribing
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: Patients age 18 years and older prescribed two or more opioids or an opioid and benzodiazepine concurrently at discharge from a hospital-based encounter (inpatient or emergency department [ED], including observation stays)
Developer Rationale: Unintentional opioid overdose fatalities have become an epidemic in the last 20 years and a major public health concern in the United States (Rudd 2016). Reducing the number of unintentional overdoses has become a priority for numerous federal organizations including the Centers for Disease Control and Prevention (CDC), the Federal Interagency Workgroup for Opioid Adverse Drug Events, and the Substance Abuse and Mental Health Services Administration.

Concurrent prescriptions of opioids or opioids and benzodiazepines places patients at a greater risk of unintentional overdose due to the increased risk of respiratory depression (Dowell 2016). An analysis of national prescribing patterns shows that more than half of patients who received an opioid prescription in 2009 had filled another opioid prescription within the previous 30 days (NIDA 2011). Studies of multiple claims and prescription databases have shown that between 5%-15% percent of patients receive concurrent opioid prescriptions and 5%-20% of patients receive concurrent opioid and benzodiazepine prescriptions across various settings (Liu 2013, Mack 2015, Park 2015). Patients who have multiple opioid prescriptions have an increased risk for overdose (Jena 2014). Rates of fatal overdose are ten times higher in patients who are co-dispensed opioid analgesics and benzodiazepines than opioids alone (Dasgupta 2015). The number of opioid overdose deaths involving benzodiazepines increased 14% on average each year from 2006 to 2011, while the number of opioid analgesic overdose deaths not involving benzodiazepines did not change significantly (Jones 2015). Furthermore, concurrent use of benzodiazepines with opioids was prevalent in 31%-51% of fatal overdoses (Dowell 2016). Emergency Department (ED) visit rates involving both opioid analgesics and benzodiazepines increased from 11.0 per 100,000 in 2004 to 34.2 per 100,000 population in 2011 (Jones 2015). One study found that eliminating concurrent use of opioids and benzodiazepines could reduce the risk of opioid overdose-related ED and inpatient visits by 15 percent and potentially could have prevented an estimated 2,630 deaths related to opioid painkiller overdoses in 2015 (Sun 2017).

A recent study on The Opioid Safety Initiative in the Veterans Health Administration (VHA), which includes the opioid and benzodiazepine concurrent prescribing measure that the Safe Use of Opioids - Concurrent Prescribing measure is based on, as well as an audit and feedback tool, was associated with a decrease of 20.67 percent overall and 0.86 percent patients per month (781 patients per month) receiving concurrent benzodiazepine with an opioid among all adult VHA patients who filled outpatient opioid prescriptions from October 2012 to September 2014 (Lin 2017).

Adopting a measure that calculates the proportion of patients with two or more opioids or opioids and benzodiazepines concurrently has the potential to reduce preventable mortality and reduce the costs associated with adverse events related to opioid use by 1) encouraging providers to identify patients with concurrent prescriptions of

opioids or opioids and benzodiazepines and 2) discouraging providers from prescribing two or more opioids or opioids and benzodiazepines concurrently.

Numerator Statement: Patients prescribed two or more opioids or an opioid and benzodiazepine at discharge.
Denominator Statement: Patients age 18 years and older prescribed an opioid or a benzodiazepine at discharge from a hospital-based encounter (inpatient stay less than or equal to 120 days or emergency department encounters, including observation stays) during the measurement period.
Denominator Exclusions: Denominator exclusions: The following encounters are excluded from the denominator:
- Encounters for patients with an active diagnosis of cancer during the encounter
- Encounters for patients who are ordered for palliative care during the encounter
- Inpatient encounters with length of stay greater than 120 days

Denominator exceptions: None.

Measure Type: Process
Data Source: Electronic Health Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A-New Measure

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this process measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- The developer provided a diagram demonstrating the link between avoiding discharging patients with concurrent prescriptions for two or more opioids or an opioid and benzodiazepine and a reduced risk of adverse drug events.
- The developer provided a clinical guideline from the CDC Guideline for Prescribing Opioids for Chronic Pain (2016):
    - Clinicians should avoid prescribing opioid pain medication and benzodiazepines concurrently whenever possible (Class III; Level of Evidence: A)
- The developer cited a systematic review of the body of the evidence on the effectiveness and risks of long-term opioid therapy. Based on CDC's GRADE criteria, the overall quality of the clinical evidence base for the effectiveness and risks of long-term opioid therapy (42 studies reviewed in total) ranged between types 3, 4, and insufficient.
- The developer also described a "rapid review" of "contextual evidence review" performed by CDC to supplement the clinical evidence review base; the quality of evidence for the studies included in the contextual evidence review (including original studies, systematic reviews, and clinical guidelines) was not rated using the GRADE criteria.

- The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the guideline.
- As summarized by the developer, the systematic review of the evidence suggests that there is an increased risk of overdose events associated with (1) opioid use and (2) co-prescription of opioids with benzodiazepines.
    - The developer stated that a "dose-dependent association" between opioid use and risk for overdose events, including death, was found consistently across two studies in the clinical evidence review and several epidemiologic studies in the contextual evidence review. Co-prescription of opioids with benzodiazepines was also found to increase risk for potentially fatal overdose in three studies included in the contextual evidence review. The studies found evidence of concurrent benzodiazepine use in 31 to 61 percent of those deceased from overdose. Finally, state-level evaluations of the effect of Prescription Drug Monitoring Programs on changes in prescribing and mortality outcomes were limited.
- However, the evidence review does not appear to address concurrent prescription of opioids.
- The developer included the potential unintended harms described in the evidence from prescribing changes: patients seeking heroin or other illicitly obtained opioids, and interference with appropriate pain treatment.

Questions for the Committee:

- *For structure, process, and intermediate outcome measures:*
    - *What is the relationship of this measure to patient outcomes?*
    - *How strong is the evidence for this relationship?*
    - *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

Process measure with systematic review (Box 3) → Summary of the QQC provided (Box 4) → Systematic review concludes moderate quality evidence.

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

1b. Gap in Care/Opportunity for Improvement and 1b. DisparitiesData
Maintenance measures – increased emphasis on gap and variation

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

- This is a new measure so performance data were available and presented for eight testing hospitals from three large tertiary health systems, in three states (TX, MI, and CT). All eight hospitals are located in urban areas, and were not-for-profit teaching hospitals. The hospitals varied in EHR systems (Cerner and Epic). The test sample from each health system included at least 50,000 encounters.
- Performance rates observed during testing aligned with those in the literature, between 5 to 15 percent of patients receiving concurrent opioid prescriptions, and 5 to 20 percent receiving concurrent opioid-benzodiazepine prescription in an inpatient or outpatient hospital setting. The developer reports that there were higher rates of concurrent prescribing in the inpatient setting compared to the ED/obs across test sites.

Disparities

- During testing, the measure performance was stratified for disparities in patient encounters, by age, sex, race, ethnicity, and primary payer.
- Additionally, there were performance gaps based on patient age, sex, race, ethnicity, and payer across test sites and by setting (inpatient vs. ED/obs).
- Across test sites, the performance rate in the inpatient setting was 18.2 percent. Older patients (65+ years) had lower performance rates than younger patients (18-64 years), male patients had worse performance rates than female patients, White patients had poorer performance rates compared to patients of other races, and non-Hispanic patients had worse performance rates than Hispanic or Latino patients. Finally, Medicare and Medicaid patients had poorer performance rates compared to patients with other types of insurance. All differences described were statistically significant ($p<.05$).

- Across test sites, the performance rate in the ED/obs was 6.1 percent. By race and ethnicity, white patients had lower performance rates than patients of other races, and non-Hispanic patients had slightly poorer performance compared to Hispanic or Latino patients. Finally, Medicare patients had the worst performance rate, while uninsured and self-pay patients had the best performance rate. All differences described were statistically significant (p<.05). There was no significant difference between performance rates by patients' age or sex.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive?*

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

Committee pre-evaluation comments

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

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.

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

Composite measures only: N/A

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

eMeasure Technical Advisor:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**

Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with measure developer indicates that the measure logic is feasible based on assessment by EHR vendors

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: Patient Safety project team staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Link A,(Project Team staff)

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 staff 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, etc.)?*
- *The staff 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)

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.

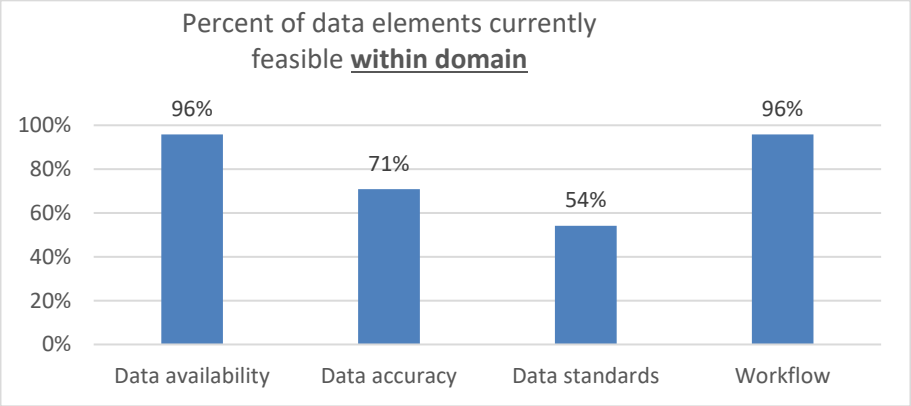
Data Specifications and Elements

- The measure is constructed using electronic health records
- All data elements are available in defined fields in electronic health records (EHRs)
- The measure developer shared the feasibility score card for this eCQM. The measure was tested in three sites using two EHR systems (Cerner and Epic):

CURRENT – SUMMARY

On a scale of 0% to 100%, how feasible is the measure currently across sites?

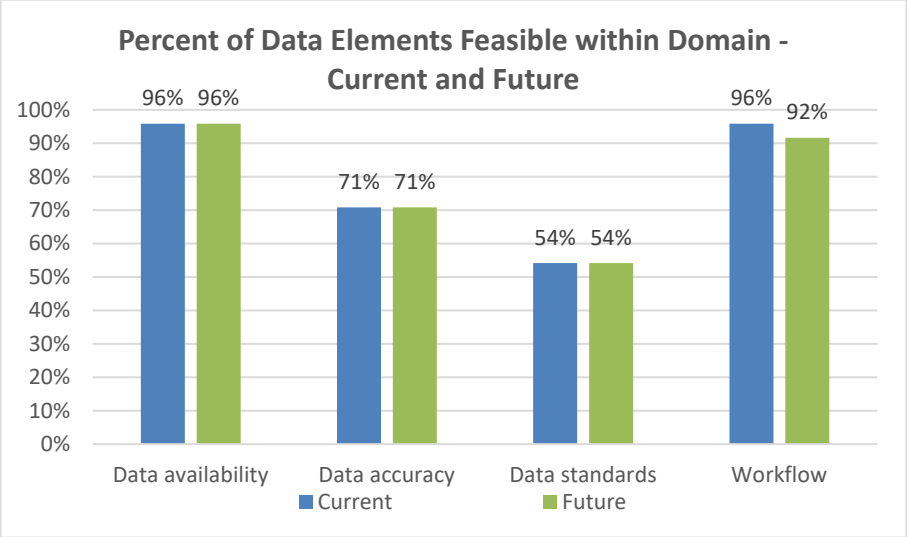
79.2%



FUTURE- SUMMARY

On a scale of 0% to 100%, how feasible is the measure in 3 to 5 years?

78.1%



Data Collection Strategy

- Value sets are housed in the Value Set Authority Center (VSAC), which has no fee for viewing/downloading.
- There are no other fees or licensing requirement to use this measure, which is in the public domain.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *If an eMeasure, does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

RATIONALE:

Committee pre-evaluation comments
Criteria 3: 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 evaluate 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**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details
The measure is not currently used in an accountability program. Per developer/steward, CMS is considering implementation plans for this measure. The measure has been submitted through the Measures Under Consideration process for the CMS Hospital Inpatient and Outpatient Quality Reporting Programs.

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 developer/steward collected feedback from clinical quality/data analytics staff, as well as other providers and some physicians at the test site locations where that participated in testing. Providers at the test sites were unsurprised by their measure performance scores, which aligned with their expectations of the rate of concurrent prescribing at their hospitals during the measurement period (October 1, 2013 - September 30, 2015).
Measure specifications were revised prior to being tested at all three test sites, and no changes were made based on discussions with providers.

Additional Feedback:
The developer/steward did not provide any further feedback.

***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**
RATIONALE:

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

4b. Usability evaluate 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 is a new measure. The developer states the findings from the field testing at the 3 health systems suggest that this measure could promote adherence to recommended clinical guidelines, improve patient care, and reduce opioid-related mortality resulting from concurrent opioids or opioid-benzodiazepines prescriptions with minimal implementation costs.

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
No unexpected findings provided by developer.

Potential harms
The developer noted that some stakeholders noted concerns about attribution of concurrent prescribing and the potential to promote abrupt cessation of medications in effort to achieve a more favorable performance score.

Other stakeholders noted the measure's potential to reduce risk of harm to patients throughout the continuum of care, adding that the decision to continue concurrent opioids and benzodiazepines until further follow-up should be made in the best interest of the patient to avoid unintended consequences.

Additional Feedback:
During discussions with the expert workgroup, experts suggested that this measure could promote better medication reconciliation practices from opioids and benzodiazepines.

In the 2016-2017 pre-rulemaking deliberations, the Measure Applications Partnership Hospital Workgroup did not support the measure for rulemaking. The measure was not supported since there are times when concurrent prescriptions of opioids and benzodiazepines are appropriate. The Workgroup was also concerned that patients may unintentionally suffer withdrawal symptoms if previously prescribed opioids and/or benzodiazepines are reduced and/or stopped prior to discharge.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

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

Committee pre-evaluation comments
Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- This measure is related to #2940: Use of Opioids at High Dosage in Persons Without Cancer, #2950: Use of Opioids from Multiple Providers in Persons Without Cancer, and #2951: Use of Opioids from Multiple Providers and at High Dosage in Persons Without Cancer.

Harmonization

- Harmonization refers to the measure specifications – not the setting or data source. The developer stated that this measure's specifications are harmonized with existing measures where possible but there are several key differences:
    - The eligible population for this measure includes not only patients prescribed at least one opioid at discharge, but also patients prescribed at least one benzodiazepine at discharge. Measures #2940, #2950 and #2951 do not include benzodiazepines in the measure focus.
    - The developer stated that Schedule II and Schedule III opioids are in scope of this measure per expert consensus. Measures #2940, #2950 and #2951 also include Schedule IV opioids.
    - This measure assesses patients across the hospital inpatients and outpatient settings. Measures #2940, #2950 and #2951 focus on the prescription drug health plan level.

Committee pre-evaluation comments
Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
    - XX support the measure
    - YY do not support the measure

Scientific Acceptability

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 subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 3316
Measure Title: Safe Use of Opioids – Concurrent Prescribing

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?
   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)
   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?
   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
   ☒Yes (go to Question #5)
   ☐No (go to Question #8)

Reliability was assessed using EHR-extracted data from each of three test sites (eight hospitals total) for the time frame October 1, 2013, to September 30, 2015. Across the three test sites, data were received for 274,499 encounters (107,020 inpatient encounters and 167,479 ED/observation encounters).

5. 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.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☒Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

The split-half approach was used to estimate the reliability of the performance rate. The measure's reliability coefficient across eight hospitals was 0.99 (95% CI: 0.98, 0.99). This result indicates that the hospital-level performance rate has excellent reliability, meaning that differences in hospital performance reflect true differences in quality as opposed to measurement error or noise.

6. **RATING (score level) -** What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☒High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as

INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☒Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

Data element validity testing evaluated whether the measure specification correctly identified all the data elements required to calculate the measure score. This method quantifies the percentage agreement, Kappa statistic, sensitivity, specificity, and negative and positive predictive values between electronically extracted EHR data and manually abstracted EHR data.

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

The Kappa values calculated through data element validity testing suggest moderate levels of agreement between the data extract generated from the EHR systems and the manually abstracted data.

11. OVERALL RELIABILITY RATING

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 (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

1. Were all potential threats to validity that are relevant to the measure empirically assessed?
   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.
   ☒Yes (go to Question #2)
   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #3)

☒No (go to Question #3)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

This is a process measure

a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

b. Are social risk factors included in risk model? ☐Yes ☐No

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☐Yes (please explain below then go to Question #4)

☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☐Yes (please explain below then go to Question #5)

☒No (go to Question #5)

- To identify statistically significant differences in performance, the developer conducted a t-test for each subgroup within each patient characteristic (age, sex, race, ethnicity, or payer)
- The results demonstrated that statistically significant differences can be detected between hospitals and between demographic characteristics (age, sex, race, ethnicity, and primary payer). The gaps in performance between hospitals and demographic groups indicate that there is room for improvement in performance rates of concurrent prescribing.
- The performance rate for the inpatient setting (n = 8) across test sites was 18.2 percent and the performance rate for the ED (n = 8) across sites was 6.1 percent. They also observed variation in performance rates across the eight hospitals with rates ranging from a low of 6.3 percent to a high of 31.3 percent in the inpatient environment and 4.6 to 8.5 percent in the ED setting.

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

☐Yes (please explain below then go to Question #6)

☐No (go to Question #6)

☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Missing data are not a threat to validity for the measure. Data elements required to calculate the performance rate are ones in which absence of data in a data field reflects the absence of a prescription at discharge. The developer did not assess the frequency of missing data because we did not find any significant issues in the extracted or abstracted data.

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☒Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

Face Validity: twelve expert work group (EWG) members and three testing site affiliated staff (N = 15 respondents) evaluated the face validity of the measure and measure score (after field testing was completed) through a survey.

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

Most respondents (73 percent) strongly agreed or agreed that the measure will likely reduce the incidence of concurrent prescribing of opioid-opioid and opioid-benzodiazepines at discharge in the inpatient and ED settings.

Table. Results of face validity evaluation

Rating	Number of EWG Members
Strongly agree	1
Agree	10
Disagree	3
Strongly disagree	1

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

☐Yes (go to Question #11)
☒No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
☐Yes (go to Question #12)
☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #14)
☐Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?
☒Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with <u>patient-level data elements</u>?
TIPS: Prior validity studies of the same data elements may be submitted
☒Yes (go to Question #15)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if <u>no</u> score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. 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.*
TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☒Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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 (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]
☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High
☐Moderate
☐Low (please explain below)
☐Insufficient (please explain below)

N/A

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Not applicable

Measure Title: Safe Use of Opioids – Concurrent Prescribing

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 11/1/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*

- *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*

- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

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

Outcome

☐ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the 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*): Click here to name the intermediate outcome

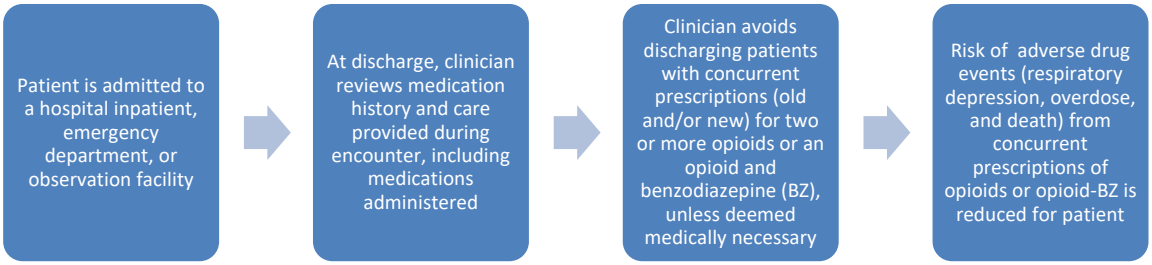
☒ Process: Patients ages 18 years and older prescribed two or more opioids or an opioid and benzodiazepine concurrently at discharge from a hospital-based encounter (inpatient or emergency department [ED] encounters, including observation stays)

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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 Safe Use of Opioids – Concurrent Prescribing measure has the potential to improve patient safety by changing clinicians' prescribing practices, specifically to reduce the rate of concurrent prescriptions of opioids or opioid-benzodiazepine at discharge from the facility.

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.)

Not applicable

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

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.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☒ Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• CDC Guideline for Prescribing Opioids for Chronic Pain (2016), Recommendation 11 • Deborah Dowell, MD, Tamara M. Haegerich, PhD, Roger Chou, MD, on behalf of CDC • Published March 18, 2016 • Dowell D, Haegerich TM, Chou R. (2016). CDC Guideline for Prescribing Opioids for Chronic Pain — United States, 2016. *Morbidity and Mortality Weekly Report, Recommendations and Reports*, *65*(1), 1–49. DOI: http://dx.doi.org/10.15585/mmwr.rr6501e1 • https://www.cdc.gov/mmwr/volumes/65/rr/pdfs/rr6501e1.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"Clinicians should avoid prescribing opioid pain medication and benzodiazepines concurrently whenever possible (recommendation category: A, evidence type: 3)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**Type 3 evidence:** Observational studies or randomized clinical trials with notable limitations.

Provide all other grades and definitions from the evidence grading system	**Type 1 evidence:** Randomized clinical trials or overwhelming evidence from observational studies. **Type 2 evidence:** Randomized clinical trials with important limitations, or exceptionally strong evidence from observational studies. **Type 4 evidence:** Clinical experience and observations, observational studies with important limitations, or randomized clinical trials with several major limitations.
Grade assigned to the **recommendation** with definition of the grade	**Category A recommendation:** Applies to all people; most patients should receive the recommended course of action.
Provide all other grades and definitions from the recommendation grading system	**Category B recommendation:** Individual decision making needed; different choices will be appropriate for different patients. Clinicians help patients arrive at a decision consistent with patients' values and preferences and specific clinical situations.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	CDC conducted a systematic review, updating an earlier AHRQ-sponsored systematic review (2014) on the effectiveness and risks of long-term opioid treatment of chronic pain with more recent publications. Based on CDC's GRADE criteria, the overall quality of the clinical evidence base for the effectiveness and risks of long-term opioid therapy (42 studies reviewed in total) ranged between types 3, 4, and insufficient. As a result, CDC conducted a "rapid review" of "contextual evidence review" to supplement the clinical evidence review base; the quality of evidence for the studies included in the contextual evidence review (including original studies, systematic reviews, and clinical guidelines, excluding grey literature) was not rated using the GRADE criteria, and should be interpreted accordingly. In addition to the clinical evidence and contextual evidence reviews, CDC solicited input from experts, federal partners, stakeholders, and the general public while developing specific recommendations for prescribing opioids for chronic pain. Studies in the clinical evidence base relevant to guideline – type of study: • Dunn (2010) – large fair-quality retrospective cohort study

	• Gomes (2011) – good-quality, population-based, nested case-control study Studies in the contextual evidence base relevant to guideline – type of study: • Bohnert (2011) – case-cohort study • Zedler (2014) – retrospective, nested, case-control analysis • Gwira Baumblatt (2014) – matched case-control study • Paulozzi (2012) – matched case-control study • Liang (2015) – longitudinal cohort study • Dasgupta (2015) – prospective observational cohort study Jones (2015) – analysis of opioid analgesic and benzodiazepine nonmedical use-related ED visits from the Drug Abuse Warning Network and drug overdose deaths from the National Vital Statistics System, 2004−2011
Estimates of benefit and consistency across studies	A "dose-dependent association" between opioid use and risk for overdose events, including death, was found consistently across two studies in the clinical evidence review and several epidemiologic studies in the contextual evidence review. Co-prescription of opioids with benzodiazepines was also found to increase risk for potentially fatal overdose in three studies included in the contextual evidence review. The studies found evidence of concurrent benzodiazepine use in 31 to 61 percent of those deceased from overdose. Finally, state-level evaluations of the effect of Prescription Drug Monitoring Programs on changes in prescribing and mortality outcomes were limited.
What harms were identified?	Potential unintended harm from prescribing changes (for example, dose reductions) include patients seeking heroin or other illicitly obtained opioids, and interference with appropriate pain treatment. CDC identified only one qualitative study reporting an association between patients receiving "an abuse-deterrent formulation of OxyContin and heroin use" and switching to another opioid. No other studies of potential negative consequences of prescribing changes were identified.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This guideline was published in 2016 and is the most recent systematic review completed. We are not aware of additional systematic reviews that have emerged since it was completed.

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1. Evidence, Performance Gap, Priority – 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

Opioids_Evidence.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.

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.

Unintentional opioid overdose fatalities have become an epidemic in the last 20 years and a major public health concern in the United States (Rudd 2016). Reducing the number of unintentional overdoses has become a priority for numerous federal organizations including the Centers for Disease Control and Prevention (CDC), the Federal Interagency Workgroup for Opioid Adverse Drug Events, and the Substance Abuse and Mental Health Services Administration.

Concurrent prescriptions of opioids or opioids and benzodiazepines places patients at a greater risk of unintentional overdose due to the increased risk of respiratory depression (Dowell 2016). An analysis of national prescribing patterns shows that more than half of patients who received an opioid prescription in 2009 had filled another opioid prescription within the previous 30 days (NIDA 2011). Studies of multiple claims and prescription databases have shown that between 5%-15% percent of patients receive concurrent opioid prescriptions and 5%-20% of patients receive concurrent opioid and benzodiazepine prescriptions across various settings (Liu 2013, Mack 2015, Park 2015). Patients who have multiple opioid prescriptions have an increased risk for overdose (Jena 2014). Rates of fatal overdose are ten times higher in patients who are co-dispensed opioid analgesics and benzodiazepines than opioids alone (Dasgupta 2015). The number of opioid overdose deaths involving benzodiazepines increased 14% on average each year from 2006 to 2011, while the number of opioid analgesic overdose deaths not involving benzodiazepines did not change significantly (Jones 2015). Furthermore, concurrent use of benzodiazepines with opioids was prevalent in 31%-51% of fatal overdoses (Dowell 2016). Emergency Department (ED) visit rates involving both opioid analgesics and benzodiazepines increased from 11.0 per 100,000 in 2004 to 34.2 per 100,000 population in 2011 (Jones 2015). One study found that eliminating concurrent use of opioids and benzodiazepines could reduce the risk of opioid overdose-related ED and inpatient visits by 15 percent and potentially could have prevented an estimated 2,630 deaths related to opioid painkiller overdoses in 2015 (Sun 2017).

A recent study on The Opioid Safety Initiative in the Veterans Health Administration (VHA), which includes the opioid and benzodiazepine concurrent prescribing measure that the Safe Use of Opioids - Concurrent Prescribing measure is based on, as well as an audit and feedback tool, was associated with a decrease of 20.67 percent overall and 0.86 percent patients per month (781 patients per month) receiving concurrent benzodiazepine with an opioid among all adult VHA patients who filled outpatient opioid prescriptions from October 2012 to September 2014 (Lin 2017).

Adopting a measure that calculates the proportion of patients with two or more opioids or opioids and benzodiazepines concurrently has the potential to reduce preventable mortality and reduce the costs associated with adverse events related to opioid use by 1) encouraging providers to identify patients with concurrent prescriptions of opioids or opioids and

benzodiazepines and 2) discouraging providers from prescribing two or more opioids or opioids and benzodiazepines concurrently.

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.*

Three large tertiary health systems, representing eight hospitals total, in three states (TX, MI, and CT) field tested the measure. All eight hospitals are located in urban areas, and are not-for-profit teaching hospitals. The hospitals varied in EHR systems (Cerner and Epic). The test sample from each health system included at least 50,000 encounters. A detailed breakdown of the characteristics of the measured facilities and the patient population can be found in sections 1.5 and 1.6 of the attached Measure Testing form.

The measure performance, including the denominator, numerator, and the performance rate by hospital and setting, is presented below. Performance rates are calculated at the encounter level, which means there could be multiple encounters for one patient.

Hospital ID 1:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 22,746
--- Numerator: 5,323
--- Performance rate: 23.40
--- 95% confidence interval: (23.38, 23.42)
- ED/observation:
--- Denominator: 13,450
--- Numerator: 735
--- Performance rate: 5.46
--- 95% confidence interval: (5.45, 5.48)

Hospital ID 2:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 4,401
--- Numerator: 353
--- Performance rate: 8.02
--- 95% confidence interval: (8, 8.05)
- ED/observation:
--- Denominator: 13,670
--- Numerator: 626
--- Performance rate: 4.58
--- 95% confidence interval: (4.57, 4.59)

Hospital ID 3:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 11,743
--- Numerator: 741
--- Performance rate: 6.31
--- 95% confidence interval: (6.3, 6.32)
- ED/observation:
--- Denominator: 17,992
--- Numerator: 918
--- Performance rate: 5.10
--- 95% confidence interval: (5.09, 5.11)

Hospital ID 4:
- Data collection period: October 1, 2013 - September 30, 2015

- Inpatient:
--- Denominator: 20,254
--- Numerator: 3,399
--- Performance rate: 16.78
--- 95% confidence interval: (16.77, 16.8)
- ED/observation:
--- Denominator: 47,081
--- Numerator: 2,377
--- Performance rate: 5.05
--- 95% confidence interval: (5.04, 5.06)

Hospital ID 5:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 9,270
--- Numerator: 1,367
--- Performance rate: 14.75
--- 95% confidence interval: (14.72, 14.77)
- ED/observation:
--- Denominator: 13,694
--- Numerator: 1,162
--- Performance rate: 8.49
--- 95% confidence interval: (8.47, 8.5)

Hospital ID 6:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 9,155
--- Numerator: 2,866
--- Performance rate: 31.31
--- 95% confidence interval: (31.27, 31.34)
- ED/observation:
--- Denominator: 10,554
--- Numerator: 866
--- Performance rate: 8.21
--- 95% confidence interval: (8.19, 8.22)

Hospital ID 7:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 9,413
--- Numerator: 1,430
--- Performance rate: 15.19
--- 95% confidence interval: (15.17, 15.21)
- ED/observation:
--- Denominator: 20,115
--- Numerator: 1,130
--- Performance rate: 5.62
--- 95% confidence interval: (5.61, 5.63)

Hospital ID 8:
- Data collection period: October 1, 2013 - September 30, 2015
- Inpatient:
--- Denominator: 20,038
--- Numerator: 4,004
--- Performance rate: 19.98
--- 95% confidence interval: (19.96, 20)

- ED/observation:
--- Denominator: 30,923
--- Numerator: 2,345
--- Performance rate: 7.58
--- 95% confidence interval: (7.57, 7.59)

Inpatient:
- Mean: 16.97
- Standard deviation: 8.09
- Minimum: 6.31
- Maximum: 31.31
- Interquartile range:
--- 10th percentile: 7.51
--- 25th percentile: 13.07
--- 50th percentile: 15.99
--- 75th percentile: 20.84
--- 90th percentile: 25.77

ED/observation:
- Mean: 6.26
- Standard deviation: 1.57
- Minimum: 4.58
- Maximum: 8.49
- Interquartile range:
--- 10th percentile: 4.91
--- 25th percentile: 5.09
--- 50th percentile: 5.54
--- 75th percentile: 7.74
--- 90th percentile: 8.29

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.
Data have been included in Section 1b.2; these data represent performance over time, from October 2013 - September 2015 data collection periods.

Overall, performance rates observed during testing aligned with those in the literature, between 5 to 15 percent of patients receiving concurrent opioid prescriptions, and 5 to 20 percent receiving concurrent opioid-benzodiazepine prescription in an inpatient or outpatient hospital setting.

Citations:

Liu, Y., Logan, J. E., Paulozzi, L. J., Zhang, K., & Jones, C. M. (2013). Potential misuse and inappropriate prescription practices involving opioid analgesics. American Journal of Managed Care, 19(8), 648–665. Retrieved [March 20, 2016] from http://www.ncbi.nlm.nih.gov/pubmed/24304213
Mack, K. A., Zhang, K., Paulozzi, L., & Jones, C. (2015). Prescription practices involving opioid analgesics among Americans with Medicaid, 2010. Journal of Health Care for the Poor and Underserved, 26(1), 182–198. Retrieved [March 20, 2016] from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4365785/
Park, T. W., Saitz, R., Ganoczy, D., Ilgen, M. A., & Bohnert, A. S. (2015). Benzodiazepine prescribing patterns and deaths from drug overdose among US veterans receiving opioid analgesics: Case-cohort study. British Medical Journal, 350:h2698. Retrieved [March 20, 2016] from http://www.bmj.com/content/bmj/350/bmj.h2698.full.pdf

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.

During testing, the measure performance was stratified for disparities, in patient encounters, by age, sex, race, ethnicity, and primary payer.

Measure Performance by Age:
- 18-64
--- Denominator: 78,450 (inpatient); 142,606 (ED/observation)
--- Numerator: 13,803 (inpatient); 8,678 (ED/observation)
--- Performance rate: 17.6% (inpatient); 6.1% (ED/observation)
- 65+
--- Denominator: 28,570 (inpatient); 24,873 (ED/observation)
--- Numerator: 5,680 (inpatient); 1,481 (ED/observation)
--- Performance rate: 19.9% (inpatient); 6.0% (ED/observation)

Measure Performance by Sex:
- Male
--- Denominator: 40,528 (inpatient); 71,207 (ED/observation)
--- Numerator: 9,223 (inpatient); 4,386 (ED/observation)
--- Performance rate: 22.8% (inpatient); 6.2% (ED/observation)
- Female
--- Denominator: 66,392 (inpatient); 96,179 (ED/observation)
--- Numerator: 10,250 (inpatient); 5,769 (ED/observation)
--- Performance rate: 15.4% (inpatient); 6.0% (ED/observation)

Measure Performance by Race:
- White
--- Denominator: 65,007 (inpatient); 88,046 (ED/observation)
--- Numerator: 12,230 (inpatient); 5,931 (ED/observation)
--- Performance rate: 18.8% (inpatient); 6.7% (ED/observation)
- Black
--- Denominator: 23,876 (inpatient); 54,495 (ED/observation)
--- Numerator: 4,232 (inpatient); 2,874 (ED/observation)
--- Performance rate: 17.7% (inpatient); 5.3% (ED/observation)
- Other
--- Denominator: 13,696 (inpatient); 19,585 (ED/observation)
--- Numerator: 2,227 (inpatient); 993 (ED/observation)
--- Performance rate: 16.3% (inpatient); 5.1% (ED/observation)

Measure Performance by Ethnicity:
- Hispanic
--- Denominator: 10,483 (inpatient); 17,316 (ED/observation)
--- Numerator: 1,679 (inpatient); 845 (ED/observation)
--- Performance rate: 16.0% (inpatient); 4.9% (ED/observation)
- Non-Hispanic
--- Denominator: 91,819 (inpatient); 144,794 (ED/observation)
--- Numerator: 17,007 (inpatient); 8,975 (ED/observation)
--- Performance rate: 18.5% (inpatient); 6.2% (ED/observation)

Measure Performance by Primary Payer:
- Medicare
--- Denominator: 32,204 (inpatient); 27,812 (ED/observation)
--- Numerator: 6,642 (inpatient); 1,888 (ED/observation)
--- Performance rate: 20.6% (inpatient); 6.8% (ED/observation)
- Medicaid
--- Denominator: 17,166 (inpatient); 38,205 (ED/observation)

--- Numerator: 2,938 (inpatient); 2,200 (ED/observation)
--- Performance rate: 17.1% (inpatient); 5.8% (ED/observation)
- Private insurance
--- Denominator: 47,894 (inpatient); 74,112 (ED/observation)
--- Numerator: 8,017 (inpatient); 4,580 (ED/observation)
--- Performance rate: 16.7% (inpatient); 6.2% (ED/observation)
- Self-pay or uninsured
--- Denominator: 4,666 (inpatient); 11,653 (ED/observation)
--- Numerator: 1,143 (inpatient); 535 (ED/observation)
--- Performance rate: 24.5% (inpatient); 4.6% (ED/observation)
- Other
--- Denominator: 1,230 (inpatient); 1,618 (ED/observation)
--- Numerator: 297 (inpatient); 113 (ED/observation)
--- Performance rate: 24.1% (inpatient); 7.0% (ED/observation)

Summary of Results:

There were higher rates of concurrent prescribing in the inpatient setting compared to the ED/obs across test sites. Additionally, there were performance gaps based on patient age, sex, race, ethnicity, and payer across test sites and by setting (inpatient vs. ED/obs).

Across test sites, the performance rate in the inpatient setting was 18.2 percent. Older patients (65+ years) had worse performance rates than younger patients (18-64 years), male patients had worse performance rates than female patients, White patients had poorer performance rates compared to patients of other races, and non-Hispanic patients had worse performance rates than Hispanic or Latino patients. Finally, Medicare and Medicaid patients had poorer performance rates compared to patients with other types of insurance. (Because the sample sizes of patients who reported having an "other" primary payer and patients who were self-pay or uninsured both were much smaller than the other payer types across test sites, their performance rates should not be considered clinically significant). All differences described were statistically significant ($p<.05$).

Across test sites, the performance rate in the ED/obs was 6.1 percent. By race and ethnicity, white patients had worse performance rates than patients of other races, and non-Hispanic patients had slightly poorer performance compared to Hispanic or Latino patients. Finally, Medicare patients had the worst performance rate, while uninsured and self-pay patients had the best performance rate. (Because the sample sizes of patients who reported having an "other" primary payer at all test sites and patients who were self-pay or uninsured at two of the three sites were much smaller than the other payer types across test sites, their performance rates should not be considered clinically significant.). All differences described were statistically significant ($p<.05$). There was no significant difference between performance rates by patients' age or sex.

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**
See response to Question 1b.4.

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.)*
No link exists, specifications are attached in accordance with question S.2a.

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:** Opioids_eCQMSpecs.zip

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 **Attachment:** Opioids_ValueSets.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.
Not applicable

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).
Patients prescribed two or more opioids or an opioid and benzodiazepine at discharge.

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).
Presence of two or more new opioids at discharge resulting in concurrent therapy is represented by QDM datatype and value set of Medication, Discharge: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2).

Presence of a new opioid and a new benzodiazepine prescription at discharge resulting in concurrent therapy is represented by QDM datatype and value sets of Medication, Discharge: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2) and Medication, Discharge: Benzodiazepines (2.16.840.1.113762.1.4.1125.1).

Presence of an existing opioid and a new opioid or benzodiazepine prescription at discharge resulting in concurrent therapy is represented by QDM datatypes and value sets of Medication, Active: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2) and Medication, Discharge: Benzodiazepines (2.16.840.1.113762.1.4.1125.1) or Medication, Discharge: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2).

Presence of an existing benzodiazepine and a new opioid prescription at discharge resulting in concurrent therapy is represented by QDM datatypes and value sets of Medication, Active: Benzodiazepines (2.16.840.1.113762.1.4.1125.1) and Medication, Discharge: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2).

Presence of an existing benzodiazepine and an existing opioid prescription at discharge resulting in concurrent therapy is represented by QDM datatype and value sets of Medication, Active: Benzodiazepines (2.16.840.1.113762.1.4.1125.1) and Medication, Active: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2).

Presence of two or more existing opioids at discharge resulting in concurrent therapy is represented by QDM datatype and value set of Medication, Active: Schedule II and Schedule III Opioids (2.16.840.1.113762.1.4.1125.2).

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)*
Patients age 18 years and older prescribed an opioid or a benzodiazepine at discharge from a hospital-based encounter (inpatient stay less than or equal to 120 days or emergency department encounters, including observation stays) during the measurement period.

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).
Inpatient Encounters are represented using the QDM datatype and value set of Encounter, Performed: Encounter Inpatient (OID: 2.16.840.1.113883.3.666.5.307). Length of stay is calculated within the measure based on encounter start and end dates. ED Encounters including observation stay are represented using the QDM datatype and value set of Encounter, Performed: Encounter ED and Observation Stay (OID: 2.16.840.1.113883.3.3157.1002.81).

Patients with an opioid or a benzodiazepine active on admission and continued at discharge are represented by the following QDM datatype and value sets:
- Medication, Active: Schedule II and Schedule III Opioids (OID: 2.16.840.1.113762.1.4.1125.2)
- Medication, Active: Benzodiazepines (OID: 2.16.840.1.113762.1.4.1125.1)

Patients who received a new opioid or benzodiazepine prescription at discharge from a qualifying encounter, not those patients who were given an opioid or benzodiazepine as part of their encounter treatment, are represented by the following QDM datatype and value sets:
- Medication, Discharge: Schedule II and Schedule III Opioids (OID: 2.16.840.1.113762.1.4.1125.2)
- Medication, Discharge: Benzodiazepines (OID: 2.16.840.1.113762.1.4.1125.1)

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/. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Denominator exclusions: The following encounters are excluded from the denominator:
- Encounters for patients with an active diagnosis of cancer during the encounter
- Encounters for patients who are ordered for palliative care during the encounter
- Inpatient encounters with length of stay greater than 120 days

Denominator exceptions: None.

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.)*
Active cancer diagnosis or palliative care order during the encounter are represented using the QDM datatype and following value sets:
- Diagnosis: Cancer (2.16.840.1.113883.3.526.3.1010)
- Intervention, Performed: Palliative care (2.16.840.1.113762.1.4.1125.3)
- Intervention, Order: Palliative care (2.16.840.1.113762.1.4.1125.3)

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.)*
Not applicable; 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.)*
Please see the attached HQMF specifications for the complete measure logic. Additionally, a flow diagram of the denominator and numerator logic is attached to the NQF submission form as a supplemental document in response to question A.1, ´Opioids_LogicFlow_for S.14 response.pdf´.

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.
Not applicable; 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.
Not applicable; 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 human readable and XML artifacts of the health quality measures format (HQMF) of the measure are contained in the eCQM specifications attached in question S.2a. 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)*
Emergency Department and Services, 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.)*
Not applicable

2. Validity – See attached Measure Testing Submission Form
Opioids_Testing.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.

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Not applicable
Measure Title: Safe Use of Opioids – Concurrent Prescribing
Date of Submission: 11/1/2017
Type of Measure:

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

Instructions

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

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who

received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

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

Not applicable. We did not use an existing data set to test this measure; instead, we partnered with hospitals to extract data from their EHR systems (described in question 1.5). In alignment with the measure specification, we asked hospitals to submit patient-level data for patients who qualify for the initial population over a two-year period, or patients 18 years and older (as of the date of the encounter) who are prescribed at least one opioid or benzodiazepine at discharge from a hospital-based encounter (inpatient stay of fewer than or equal to 120 days or ED, including hospital observation stays). The measure calculates the proportion of patients ages 18 years and older prescribed two or more opioids, or an opioid and a benzodiazepine at discharge from a hospital-based encounter. Encounters from inpatient psychiatric, hospice or palliative care, substance abuse or mental health services, dialysis, ancillary care settings, or ambulatory surgical centers are excluded. The measure also excludes encounters in which there is an order for palliative care or a diagnosis of cancer.

1.3. What are the dates of the data used in testing? October 1, 2013 – September 30, 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician

☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

Three large tertiary health systems, representing eight hospitals total, in three states (Connecticut, Michigan, and Texas) field tested the measure. All eight hospitals are located in urban areas and are not-for-profit teaching hospitals. The hospitals varied in EHR systems (Cerner and Epic). Table 1 breaks down the characteristics of the participating hospitals included in the field testing of the measure.

Table 1. Field testing hospital characteristics

Hospital System	Hospital ID	State	# of Beds	Teaching Status	EHR System	Inception of Current EHR System
Test Site 1	1	TX	861	Teaching	Cerner	2006
	2	TX	351	Teaching	Cerner	2006
	3	TX	208	Teaching	Cerner	2006
Test Site 2	4	MI	877	Teaching	Epic	2014
	5	MI	361	Teaching	Epic	2014
	6	MI	360	Teaching	Epic	2014
	7	MI	191	Teaching	Epic	2014
Test Site 3	8	CT	944	Teaching	Epic	2013

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*)

Across the three test sites, we received data for 274,499 encounters (107,020 inpatient encounters and 167,479 ED/observation encounters). In the inpatient setting, the average age of patients was 50.1 years, ranging from a low of 47.5 years at Test Site 1 to a high of 52.8 years at Test Site 2. At Test Site 1, 57.7 percent of patients were white and 16.6 percent were black. At Test Sites 2 and 3, 63.4 and 61.9 percent of patients were white, and 26.1 and 18.6 percent were black, respectively. The vast majority of patients across test sites were non-Hispanic. The most common payer across sites was private insurance (47.2 percent) followed by Medicare (28.1 percent).

In the ED setting, the average age of patients was 45.7 years. At Test Site 1, 59.1 percent of patients were white and 21.9 percent of patients were black. At Test Sites 2 and 3, 50.8 and 52.7 percent of patients were white and 37.9 and 25.2 percent were black, respectively. The vast majority of patients across test site were non-Hispanic. The most common payers among sites varied. The proportion of patients with private insurance across sites was

47.5 percent. Tables 2 and 3 break down these demographic characteristics by setting (inpatient and ED/observation) by test site.

Table 2. Demographic characteristics of the field-testing sample in the inpatient setting

	Test Site 1		Test Site 2		Test Site 3		Across Sites (pooled data)	
	N	%	N	%	N	%	N	%
Unique patients	35,212		38,719		16,077		90,008	
Average age	47.5		52.8		49.5		50.1	
Sex								
Male	13,469	38.3%	13,531	34.90%	5,974	37.2%	32,974	36.6%
Female	21,650	61.5%	25,187	65.10%	10,103	62.8%	56,940	63.3%
Race								
White	20,324	57.7%	24,545	63.40%	9,955	61.9%	54,824	60.9%
Black	5,851	16.6%	10,096	26.10%	2,997	18.6%	18,944	21.0%
Other	7,589	21.6%	2,013	5.2%	2,645	16.5%	12,262	13.6%
Ethnicity								
Hispanic	5,587	15.9%	1,424	3.7%	2,307	14.3%	9,318	10.4%
Non-Hispanic	27,975	79.4%	34,885	90.1%	13,619	84.7%	76,479	85.0%
Payer source								
Medicare	8,175	23.2%	12,686	32.8%	4,387	27.3%	25,248	28.1%
Medicaid	4,569	13.0%	5,576	14.4%	4,049	25.2%	14,194	15.8%
Private insurance	17,529	49.8%	19,759	51.0%	5,162	32.1%	42,450	47.2%
Self-pay or uninsured	3,628	10.3%	0	0%	742	4.6%	4,370	4.9%
Others	723	2.1%	209	0.5%	208	1.3%	1,140	1.3%

Note: Table does not include patients with missing or unknown data.

Table 3. Demographic characteristics of the field-testing sample in the ED setting

	Test Site 1		Test Site 2		Test Site 3		Across Sites (pooled data)	
Demographic	N	%	N	%	N	%	N	%

Number of unique patients	38,681		72,585		24,376		135,642	
Average age	47.2		45.2		44.7		45.7	
Sex								
Male	16,730	43.3%	30,887	42.6%	11,178	45.9%	58,795	43.3%
Female	21,868	56.5%	41,695	57.4%	13,198	54.1%	76,761	56.6%
Race								
White	22,861	59.1%	36,867	50.8%	12,855	52.7%	72,583	53.5%
Black	8,455	21.9%	27,477	37.9%	6,144	25.2%	42,076	31.0%
Other	6,240	16.1%	5,487	7.6%	4,598	18.9%	16,325	12.0%
Ethnicity								
Hispanic	6,909	17.9%	2,946	4.1%	4,517	18.5%	14,372	10.6%
Non-Hispanic	29,879	77.2%	67,162	92.5%	19,587	80.4%	116,628	86.0%
Payer source								
Medicare	5,882	15.2%	12,789	17.6%	3,957	16.2%	22,628	16.7%
Medicaid	3,491	9.0%	15,980	22.0%	8,912	36.6%	28,383	20.9%
Private insurance	18,454	47.7%	37,252	51.3%	8,714	35.7%	64,420	47.5%
Self-pay or uninsured	10,136	26.2%	0	0%	65	0.3%	10,201	7.5%
Others	768	2.0%	361	0.5%	248	1.0%	1,377	1.0%

Note: Table does not include patients with missing or unknown data.

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.

- **Reliability:** To assess reliability, we used EHR-extracted data from each of three test sites (eight hospitals total) for the time frame October 1, 2013, to September 30, 2015.
- **Validity:** To assess criterion (data element) validity, we randomly selected a subset of 218 patient encounters (from the full EHR extract) in two of the three test sites. For these cases, trained abstractors manually abstracted data elements necessary for the measure calculation from each site's EHR. We then compared the manually and electronically abstracted data to assess data element validity.
- **Face validity:** We solicited feedback on the measure's face validity from clinicians, information technology professionals, and experts via a web-based survey.
- **Exclusions:** To assess the prevalence and impact of exclusion criteria, we used the same EHR-extracted data used for reliability analysis.
- **Risk adjustment:** Not applicable; this measure is not risk adjusted.
- **Meaningful difference in performance:** To assess whether meaningful differences in performance exist between patient subgroups, we used the same EHR-extracted data used for reliability analysis.

- **Missing data or bias:** Not applicable; missing data are not a threat to validity for the measure. Data elements required to calculate the performance rates are ones in which absence of data in a data field reflects the absence of a prescription at discharge. The measure calculation requires the encounter type to assess medications active at discharge. Date of birth is also required because it applies for patients 18 years of age and older. Rates of missing data on these items were negligible.

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 1.6, we collected information on the following variables using data extracted from hospital EHR systems: age, sex, race, ethnicity, and payer. This measure is based on a process that should be carried out for all patients (except those excluded), so no adjustment for patient mix is necessary. We collected information about these five variables and assessed disparities in performance rates for each group. Section 2b4 describes those results.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

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

We estimated the reliability of the performance rate using a split-half approach, which assesses the precision of the measure by characterizing the correlation of estimated measure results between two non-overlapping data sets using data across settings from all eight hospitals. Specifically, split-half correlation first takes a random sample of half of the population for each hospital. Then, it calculates the correlation of the hospital measure results between the two random halves. Repeating the randomization 2,500 times enables us to calculate an estimate of the 95 percent confidence interval for the facility-level reliability score. The higher the correlation, the higher the statistical reliability of the measures. Stated another way, the higher the correlation, the greater the amount of variation that can be explained through systematic differences across the test sites as opposed to random error (for example, sampling variation within measured entities). A reliability estimate of 0.7 has been used to define good reliability and the threshold at which meaningful differences in performance can be detected.

[Reference: Adams, John L. "The Reliability of Provider Profiling: A Tutorial." Santa Monica, CA: RAND Corporation, 2009.]

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*)

Reliability tests were conducted across hospitals to generate a reliability score for the measure. Because we are looking at measure-level reliability, the measure has one reliability score:

Table 4. Reliability score

Measure Name	Reliability Score	95% Confidence Interval
Safe Use of Opioids – Concurrent Prescribing	0.99	0.98–0.99

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The measure's reliability coefficient across eight hospitals was 0.99 (95% CI: 0.98, 0.99). This result indicates that the hospital-level performance rate has excellent reliability, meaning that differences in hospital performance reflect true differences in quality as opposed to measurement error or noise. Reliability coefficients of .90 or above reflect excellent precision in performance scores.

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 (Criterion) Validity

Data element validity testing evaluated whether the measure specification correctly identified all the data elements required to calculate the measure score. This method quantifies the percentage agreement, Kappa statistic, sensitivity, specificity, and negative and positive predictive values between electronically extracted EHR data (using a file layout that defines the data elements collected through electronic query) and manually abstracted EHR data (which use the entire record, including free-text notes fields). Each of these statistics illustrates the closeness between data element results from the two sources. In general, the higher the Kappa value, the greater the chance-adjusted agreement between the data from the two sources.

Trained abstractors performed the manual abstraction using a standardized web-based tool. The manually abstracted data represent the gold standard against which we assessed the validity of the EHR-extracted data.

Face Validity

Twelve expert work group (EWG) members and three testing site affiliated staff (N = 15 respondents) evaluated the face validity of the measure and measure score after field testing was completed). The evaluation of face validity was conducted through a survey that asked respondents whether the measure will likely reduce the incidence of concurrent prescribing of multiple opioids or opioid-benzodiazepines at discharge in the inpatient and ED settings (including hospital observation stays). For each item, respondents indicated the extent to which they agreed (1 = Strongly Disagree; 2 = Disagree, 3 = Agree; 4 = Strongly Agree). We also asked if the measure score would vary based on provider performance. We evaluated the number of respondents who strongly agreed or agreed with statements in the survey. We also reviewed respondents' comments written in the free-text portions of the survey. A list of organizations represented by respondents in the face validity survey follows:

- Winchester Medical Center, Winchester, VA
- Test Site 3 School of Medicine, New Haven, CT
- Centers for Disease Control and Prevention (CDC), Atlanta, GA
- ASC Quality Collaboration, St. Pete Beach, FL
- Albany Stratton VA Medical Center, Albany, NY
- Inflexxion, Newton, MA
- Denver Health, Denver, CO
- Texas Tech University, Lubbock, TX
- Kaiser Permanente Oakland Medical Center, Oakland, CA
- U.S. Department of Veterans Affairs, Washington, DC
- Boston University Medical Center, Boston, MA
- Performance Measurement and Strategic Alliances, PQA
- Clinical quality staff from Test Sites 1 and 2

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

Data Element Validity

We measured percent agreement, defined as the number of patients for whom both data sources, electronically and manually abstracted EHR data, agree on the presence or absence of a condition. We also used Cohen's Kappa statistic to adjust percent agreement to account for chance agreement. The Kappa score can range from -1.00 to 1.00. Although higher Kappa scores indicate higher agreement between two data sources, a low Kappa score does not necessarily represent low agreement when the data are imbalanced.

We found high levels of percent agreement between the electronically and manually abstracted data for the denominator, denominator exclusions, and numerator, as seen in Table 5.

Table 5. Agreement statistics for random sample data between EHR extraction and manual chart abstraction (n = 218)

Measure Component	Agreement (%)	Kappa	Sensitivity	Specificity
Initial population	0.85	0*	0.85	NaN
Denominator	0.85	0*	0.85	NaN
Denominator exclusions				
Palliative care	0.99	0*	0	1
Cancer	0.94	0.58	0.43	1

Numerator	0.84	0.49	0.95	0.48

Source: Data from 10/1/2013 to 9/30/2015 for two sites and 10/1/2014 to 9/30/2015 for one site.

Notes: NaN: Not calculable because the denominator in the equation is equal to zero.

*All 218 cases were contained within the denominator from the EHR. The Kappa statistic treats the 218 yes-yes agreement largely as chance agreement and penalizes this condition when applying the chance correction.

[Reference: Viera, Anthony J., and Joanne M. Garrett. "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine*, vol. 37, no.5, 2005, pp. 360–363.]

Face Validity

Twelve expert work group (EWG) members and three testing site affiliated providers (N = 15 respondents) evaluated the face validity of the measure. Table 6 presents the results of their rating of face validity.

Table 6. Results of face validity evaluation

Rating	Number of EWG Members
Strongly agree	1
Agree	10
Disagree	3
Strongly disagree	1

Most respondents (73 percent) strongly agreed or agreed that the measure will likely reduce the incidence of concurrent prescribing of opioid-opioid and opioid-benzodiazepines at discharge in the inpatient and ED settings.

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

The Kappa values calculated through data element validity testing suggest moderate levels of agreement between the data extract generated from the EHR systems and the manually abstracted data. The overall sample showed 84 percent agreement or higher for all data elements. In addition, agreement was almost perfect for two of the exclusionary data elements (palliative care and cancer). Kappa values (chance-adjusted agreement) for the denominator exclusion of cancer patients and the numerator across settings and sites had moderate agreement, at 0.58 and 0.49 for inpatient and ED/observation, respectively. The denominator, initial population, and palliative care exclusions have Kappa values of 0 because most of the extracted and abstracted data matched completely. Therefore, the Kappa statistic interprets this agreement as purely by chance, resulting in an artificially low agreement estimate.

The data elements for the denominator, initial population, and numerator had high sensitivity at both sites (ranging from 0.87 to 1.00), indicating that the structured EHR data from Test Sites 1 and 2 correctly identified patients with at least one opioid or benzodiazepine prescribed at discharge. However, sensitivity was low for the cancer and palliative care exclusions. This is because there were only a few patients in the initial population with cancer or receiving palliative care. Among those with cancer or receiving palliative care, very few were correctly identified in the electronic extract.

In addition, face validity appears to be high: 72 percent of respondents also believe that inpatient and ED settings in which there are low rates of concurrent prescribing of opioid-opioid or opioid-benzodiazpine at

discharge should score well on this measure. One EWG member commented that the measure has the potential to influence physicians' practice to meet a measure versus good patient care. Those who disagreed with the statement that the measure will likely reduce the incidence of concurrent prescribing of opioid-opioid and opioid-benzodiazepine at discharge in the inpatient and ED settings mentioned that the measure does not take into account clinically appropriate scenarios in which concurrent prescribing is necessary, unfairly penalizes hospitals or facilities with a greater proportion of chronic pain patients as delivering low-quality care, and might potentially contribute to undertreatment of pain.

2b2. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b3*

The following 2 exclusions apply to the measure:
Exclusion 1: Patients with a cancer diagnosis
Exclusion 2: Patients receiving palliative care

Rationale:
The measure is based on the 2016 CDC Guidelines for Prescribing Opioids for Chronic Pain; those guidelines recommend prescribing opioid pain medications for chronic pain outside of active cancer treatment, palliative care, and end-of-life care populations because of the unique therapeutic goals, ethical considerations, opportunities for medical supervision, and balance of risks and benefits required for opioid therapy for these patient populations. All stakeholders that provided feedback on the measure agreed the current exclusions were appropriate.

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*)

We calculated and compared the performance rates with and without each exclusion to examine the effects.

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*)

Table 7. Exclusion of patients with cancer diagnosis or receiving palliative care in inpatient settings

				Performance Rate Without Specific Denominator Exclusions		
	Number Excluded (%)	Performance Rate	Performance Rate Without Denominator Exclusions	Palliative Care	Cancer Exclusion (patients who are not in remission)	Cancer Exclusion (patients in remission only)
Total	8,726 (8.8%)	18.2%	18.5%	17.4%	17.9%	18.2%
Test Site 1	15 (0.04%)	16.4%	15.5%	16.4%	15.5%	16.4%

Test Site 2	3,982 (9.3%)	18.8%	18.0%	18.9%	17.9%	18.8%
Test Site 3	4,729 (22.7%)	20.0%	23.7%	20.7%	21.4%	20.0%

Table 8. Exclusion of patients with cancer diagnosis or receiving palliative care in ED setting

				Performance Rate Without Specific Denominator Exclusions		
	Number Excluded (%)	Performance Rate	Performance Rate Without Denominator Exclusions	Palliative Care	Cancer Exclusion (patients who are not in remission)	Cancer Exclusion (patients in remission only)
Total	3,488 (2.5%)	6.1%	6.2%	6.1%	6.1%	6.1%
Test Site 1	30 (0.07%)	5.1%	5.0%	5.1%	5.0%	5.1%
Test Site 2	408 (0.6%)	6.1%	6.0%	6.1%	6.0%	6.1%
Test Site 3	3,050 (11.1%)	7.6%	7.9%	7.6%	7.9%	7.6%

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*)

Performance rates vary little regardless of applying the denominator exclusions across sites and settings. In the inpatient setting, when including all patients 18 years of age and older who received palliative care or were diagnosed with cancer, whom the measure excludes, the performance rate increases from 18.2 (measure as specified) to 18.5 percent. Similarly, the performance rate of the ED as currently specified increases from 6.1 to 6.2 percent when including patients who meet the denominator exclusion criteria. This suggests that it is unlikely that the exclusions will put any specific test site at an advantage or disadvantage. However, for face validity, clinicians' acceptance of the measure, and consistency with clinical guidelines, we recommend that the measure exclusions remain as specified.

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** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

Not applicable.

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

Not applicable.

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

Not applicable.

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

☐ **Published literature**
☐ **Internal data analysis**
☒ **Other (please describe)** Not applicable.

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

Not applicable.

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

Not applicable.

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

Not applicable.

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*)**:** Not applicable.

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): Not applicable.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves: Not applicable.

2b3.9. Results of Risk Stratification Analysis: Not applicable.

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

Not applicable.

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

Not applicable.

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

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

We analyzed the data to determine if there were statistically significant differences in performance by hospital or by age, sex, race, ethnicity, or payer. To identify statistically significant differences in performance, we conducted a t-test for each subgroup within each patient characteristic. If the 95 percent confidence intervals did not overlap across subgroups of a patient characteristic (for example, between males and females), the difference in performance scores across these disparity groups was considered statistically significant. Otherwise, such differences were not considered statistically significant.

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*)

Performance by Hospital

Performance rates varied between the inpatient and ED settings, as well as by hospital, indicating sizable performance gaps and the importance of the measure. There were higher rates of concurrent prescribing in the inpatient setting compared with the ED/observation across sites. The performance rate for the inpatient setting (n = 8) across test sites was 18.2 percent and the performance rate for the ED (n = 8) across sites was 6.1 percent. We also observed variation in performance rates across the eight hospitals with rates ranging from a low of 6.3 percent to a high of 31.3 percent in the inpatient environment and 4.6 to 8.5 percent in the ED setting. In Table 9, as well as in Figures 1 and 2, we provide performance rates for each hospital across the three

test sites. Performance rates are calculated at the encounter level, which means there could be multiple encounters for one patient.

Table 9. Nonstratified performance rates

Hospital	Inpatient (%)	ED/Observation (%)
1	23.4	5.5
2	8.0	4.6
3	6.3	5.1
4	16.8	5.1
5	14.8	8.5
6	31.3	8.2
7	15.2	5.6
8	19.9	7.6
Mean	18.2	6.1

Figure 1. Distribution of performance rates, by hospital (inpatient)

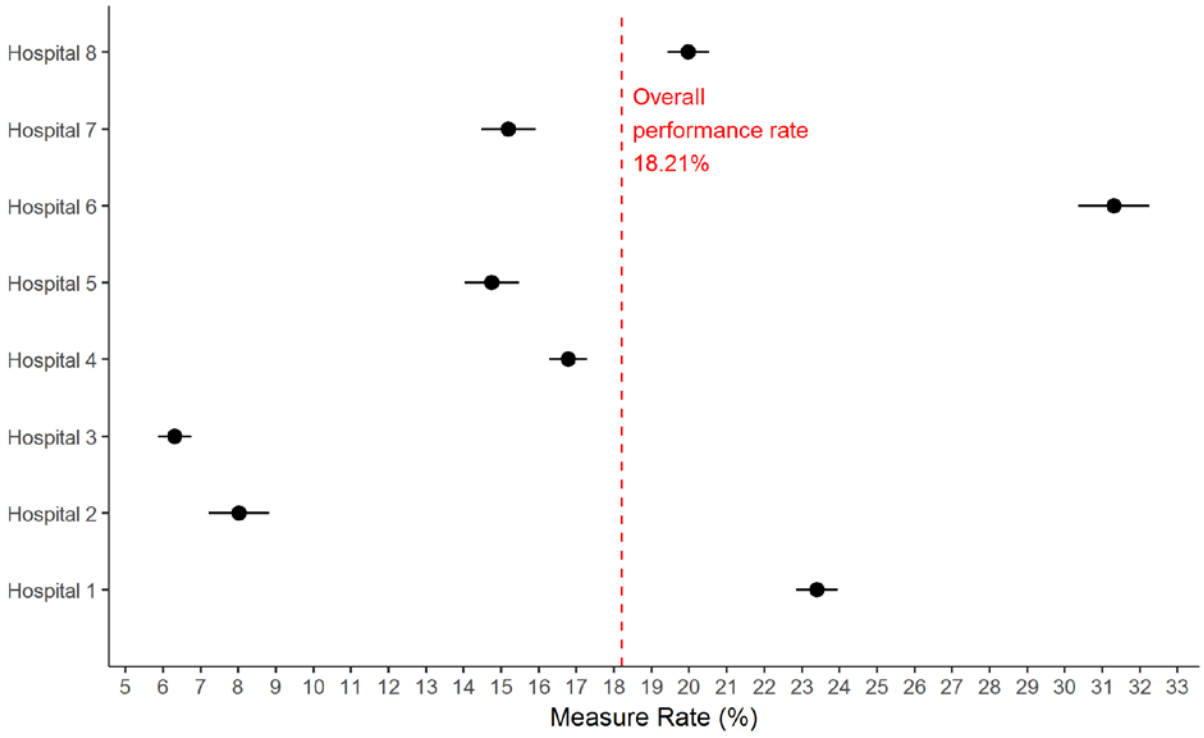
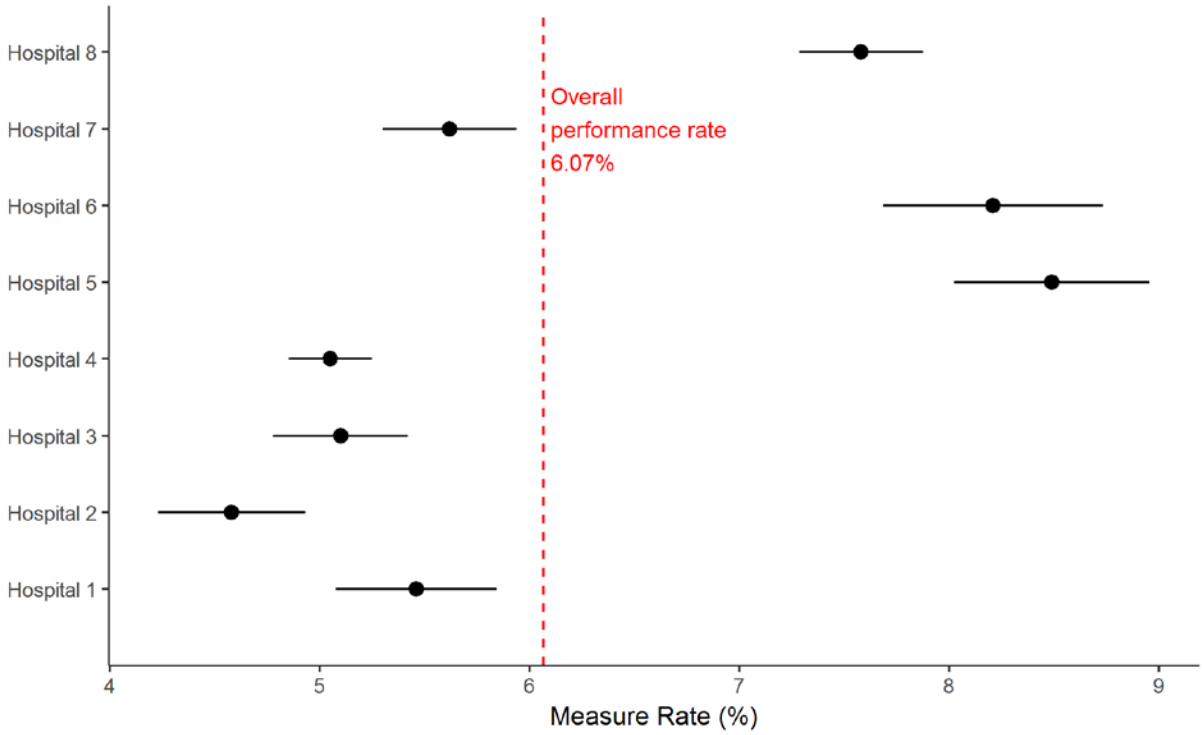


Figure 2. Distribution of performance rates, by hospital (ED/observation)



Performance by Disparity Group

We also identified performance gaps based on patients' age, sex, race and payer across test sites and by setting—all differences described below were statistically significant ($p < .05$).

Inpatient setting. The performance rate in the inpatient setting was 18.2 percent. Older patients had worse performance rates than younger patients. The performance rate for patients ages 18 to 64 was 17.6 percent and 19.9 percent for patients ages 65 and older. Male patients had worse performance rates than female patients, 22.8 and 15.4 percent, respectively. In two of the three sites, white patients had poorer performance rates compared with patients of other races, and non-Hispanic patients had worse performance rates than Hispanic or Latino patients. Medicare and Medicaid patients had poorer performance rates (20.6 and 17.1 percent, respectively) compared with patients with private insurance (16.7 percent).

ED and observation stays. The performance rate in the ED/observation was 6.1 percent. No significant gap was observed in performance by age or gender in ED/observation. Similar to findings in the inpatient setting, white patients had worse performance rates than patients of other races, and non-Hispanic patients had slightly poorer performance compared with Hispanic or Latino patients. Patients with Medicare and private insurance had worse performance rates (6.8 and 6.2 percent, respectively) than patients with Medicaid (5.8 percent), while uninsured and self-pay patients had the best performance rate at 4.6 percent.

Table 10. Performance rate (%) by patient characteristic in the Inpatient and ED settings

	Inpatient				ED			
Characteristics	**Test Site 1**	**Test Site 2**	**Test Site 3**	**Across sites**	**Test Site 1**	**Test Site 2**	**Test Site 3**	**Across sites**
Total	16.5	18.8	20.0	18.2	5.1	6.1	7.6	6.1
Age								
18 to 64	16.4	17.9	19.4	17.6	5.1	6.1	7.3	6.1

65 and older	16.8	20.9	21.8	19.9	5.0	5.6	9.5	6.0
Sex								
Male	22.5	22.3	24.4	22.8	5.5	6.1	7.4	6.2
Female	12.7	16.8	17.2	15.4	4.8	6.0	7.8	6.0
Race								
White	15.4	20.0	22.1	18.8	5.3	6.7	9.4	6.7
Black	17.7	17.1	19.5	17.7	4.8	5.3	5.7	5.3
Other	18.5	12.8	12.7	16.3	4.5	5.5	5.2	5.1
Ethnicity								
Hispanic	18.1	11.8	14.1	16.0	4.0	5.8	5.4	4.9
Non-Hispanic	16.3	19.1	21	18.5	5.2	6.0	8.1	6.2
Payer source								
Medicare	16.2	21.6	24.6	20.6	5.1	6.2	10.8	6.8
Medicaid	14.9	14.7	22.2	17.1	4.8	5.3	6.9	5.8
Private insurance	14.7	18.1	18.2	16.7	5.3	6.3	7.7	6.2
Self-pay or uninsured	26.3	NA	15.7	24.5	4.5	NA	11.9	4.6
Others*	25.5	19.9	23.5	24.1	6.6	6.3	9.4	7.0

Source: All test site data from October 1, 2013, to September 30, 2015.

Note: This table does not include patients with missing or unknown characteristics data. All comparisons between patient characteristic groups (such as ages 18 to 64 versus ages 65 and older, or male versus female) within test site, are statistically significant at the $p < .05$ level for the inpatient setting. In the ED/observation, no significant gap was observed in performance by age or gender.

*Includes all possible payers other than those listed, such as government plans (for example, federal, state, and local) that are not Medicare or Medicaid.

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?*)

The results demonstrated that statistically significant differences can be detected between hospitals and between demographic characteristics (age, sex, race, ethnicity, and primary payer). The gaps in performance between hospitals and demographic groups indicate that there is room for improvement in performance rates of concurrent prescribing.

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*)

Not applicable.

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

Not applicable.

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

Not applicable.

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Missing data are not a threat to validity for the measure. Data elements required to calculate the performance rate are ones in which absence of data in a data field reflects the absence of a prescription at discharge. For example, if data are missing from medication fields (for example, medication name), we interpret this to mean that the patient was not prescribed any medication at discharge, not that the patient was prescribed medication at discharge but this information is missing. Encounter type and discharge date are required for the measure calculation to assess medications that are active at discharge in the inpatient setting and ED/observation. Date of birth is also required, as it applies for patients ages 18 years and older. Rates of missing data on these items were negligible. We did not assess the frequency of missing data because we did not find any significant issues in the extracted or abstracted data.

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)

See response for 2b6.1

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

See response for 2b6.1

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).

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: Opioids_Feasibility.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.
Not applicable

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)*.
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.

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).

There are no other fees or licensing requirements to use this measure, which is in the public domain.

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)
Public Reporting Payment Program Regulatory and Accreditation Programs Quality Improvement (external benchmarking to organizations)	

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

Not applicable; the measure is under initial endorsement review and is not currently used in an 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?*)
CMS is considering implementation plans for this measure. There are no identified barriers to implementation in a public reporting or accountability application.

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.*)
The measure has been submitted through the Measures Under Consideration process for the CMS Hospital Inpatient and Outpatient Quality Reporting Programs.

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.
Performance results and other relevant data were compiled and presented through Powerpoint slides to the clinical quality and data analytics staff (7 people total) at the three test sites that participated in testing. These slides were then circulated to other providers in the test site who had been involved in testing or expressed an interest to see the measure's findings. At one test site, these findings were presented during a site visit to clinical quality staff and physicians.

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.**
After testing was completed, we created Powerpoint slides for each health system that included information on the following: patient characteristics of the numerator population identified in the EHR data (including average age, gender, race, and payer type); feasibility findings related to workflow; data availability, data accuracy, and data standard for both the inpatient and ED settings; tables with performance rates by setting for the test site and across test sites; number of patients in the initial population by setting; number of patients who met exclusion criteria by setting; and, the reliability estimate for each setting and across settings. For two test sites, we also presented tables showing performance rates by age, sex, ethnicity, race, and payer for the inpatient and ED settings. We scheduled separate phone conferences with each test site to present and explain these findings, including interpretation of the performance rates displayed in the tables, and discussed the measure's general feasibility, validity, reliability, and usability findings across test sites.

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.
We collected feedback from the clinical quality and data analytics staff during the presentation of measure performance and results for each test site. The materials for this presentation were then circulated to other providers in the test site who had been involved in testing or expressed an interest to see the measure's findings. For two test sites, we also received feedback from physicians whom we had interviewed during testing.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, clinical quality and data analytics staff and other providers at the test sites were unsurprised by their measure performance scores, which aligned with their expectations of the rate of concurrent prescribing at their hospitals during the measurement period (October 1, 2013 - September 30, 2015). However, physicians across test sites mentioned that they would expect the performance rates of the inpatient and ED settings to be lower using EHR data beyond September 30, 2015; as media attention surrounding the opioid epidemic increased in the past two years, test sites have implemented policies to discourage opioid prescribing at hospitals. For one health system, physicians are required, by state law, to check the state prescription drug monitoring database. One physician commented that this may have had an effect on decreasing concurrent prescribing since its enforcement in 2016.

4a2.2.3. **Summarize the feedback obtained from other users**
Not applicable

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.**
Not applicable. Measure specifications were revised prior to being tested at all three test sites, and no changes were made based on discussions with providers.

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.
Adopting a measure that calculates the proportion of patients with two or more opioids or opioids and benzodiazepines concurrently has the potential to reduce preventable mortality and reduce the costs associated with adverse events related to opioid use by 1) encouraging providers to identify patients with concurrent prescriptions of opioids or opioids and benzodiazepines and 2) discouraging providers from prescribing two or more opioids or opioids and benzodiazepines concurrently.

Our findings from field testing the measure at three health systems suggest that this measure could promote adherence to recommended clinical guidelines, improve patient care, and reduce opioid-related mortality resulting from concurrent opioids or opioid-benzodiazepines prescriptions with minimal implementation costs.

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.**
Overall, stakeholders agreed that co-prescribing of opioids or opioid-benzodiazepines is an important concept to measure. Some stakeholders noted concerns about attribution of concurrent prescribing and the potential to promote abrupt cessation of medications in effort to achieve a more favorable performance score; however, others noted the measure's potential to reduce risk of harm to patients throughout the continuum of care, adding that the decision to continue concurrent opioids and benzodiazepines until further follow-up should be made in the best interest of the patient to avoid unintended consequences.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
During discussions with the expert workgroup, experts suggested that this measure could promote better medication reconciliation practices from opioids and benzodiazepines.

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)**
2940 : Use of Opioids at High Dosage in Persons Without Cancer
2950 : Use of Opioids from Multiple Providers in Persons Without Cancer
2951 : Use of Opioids from Multiple Providers and at High Dosage in Persons Without Cancer

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

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.**
This proposed measure is a new measure. The list of Schedule II and III opioids and denominator exclusions are harmonized, where feasible, with NQF-endorsed PQA measures 2940, 2950, and 2951. The measure specifications of the proposed measure are not completely harmonized with these PQA measures as they do not include benzodiazepines in the measure focus. Below we describe the differences between the proposed measure and NQF #2940, #2950, and #2951: The eligible population for the Concurrent Prescribing measure captures not only patients prescribed at least one opioid at discharge, but also patients prescribed at least one benzodiazepine at discharge per the measure focus. Experts stressed the importance of including both opioids and benzodiazepines in the denominator to ensure that the measure takes into consideration any iatrogenic risk from co-prescribing for both populations already on opioids or benzodiazepines; Only Schedule II and Schedule III opioids are in scope of the Concurrent Prescribing measure per expert consensus. The PQA measures also include Schedule IV opioids; The Concurrent Prescribing measure assesses patients across the hospital inpatients and outpatient settings (ED, including observation stays) per the programs in which the measure will be proposed for implementation. The PQA measure focuses on the prescription drug health plan level.

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.)
Not applicable

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.

Attachment **Attachment:** Opioids_LogicFlow_for_S.14_response.pdf

Contact Information

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

Co.2 **Point of Contact:** Joseph, Clift, joseph.clift@cms.hhs.gov, 410-786-4165-
Co.3 **Measure Developer if different from Measure Steward:** Mathematica Policy Research
Co.4 **Point of Contact:** Brenna, Rabel, brabel@mathematica-mpr.com, 609-945-6564-

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.
Safe Use of Opioids Expert Workgroup:
This panel provided expertise in ED and inpatient opioid and benzodiazepine-related care and provided feedback on the measure specifications and testing results.
- Thomas Barber, M.D., Kaiser Permanente Oakland Medical Center, Oakland, CA
- Stephen Butler, Ph.D., Inflexxion, Newton, MA
- Stephen Cantrill, M.D., F.A.C.E.P., Denver Health, Denver, CO
- Deborah Dowell, M.D., M.P.H., Centers for Disease Control and Prevention (CDC), Atlanta, GA
- Thomas Emmendorfer, Pharm.D., Department of Veterans Affairs, Washington, DC
- Jeffrey Fudin, Pharm.D., D.A.A.P.M., F.C.C.P., Albany Stratton VA Medical Center, Albany, NY
- Traci Green, Ph.D., Ms.C., Boston University Medical Center, Boston, MA
- Robert Kerns, Ph.D., Yale School of Medicine, New Haven, CT
- Susan McBride, Ph.D., R.N., B.S.N., Texas Tech University, Lubbock, TX
- Deb Saine, M.S., F.A.S.H.P., Winchester Medical Center, Winchester, VA
- Donna Slosburg, B.S.N., L.H.R.M., C.A.S.C., ASC Quality Collaboration, St. Pete Beach, FL

Hospital-MDM Project Technical Expert Panel:
This panel provided overall guidance on measure development and project direction, including review of the measure specification and testing results.
- Maureen Dailey, PhD, RN, Senior Policy Fellow, American Nurses Association
- Stephen Edge, MD, Surgical Oncologist; Director, Baptist Cancer Center
- Nancy Foster, Vice President for Quality and Patient Safety Policy, American Hospital Association
- Nathan Goldstein, MD, Associate Professor, Brookdale Dept of Geriatrics and Palliative Medicine, Mount Sinai School of Medicine
- John Hertig, PharmD, MS, Pharmacist and Associate Director, Center for Medication Safety Advancement, Purdue University
- Michael Howell, MD, MPH, Associate Chief Medical Officer for Clinical Quality, University of Chicago Medicine
- Thomas Louis, PhD, Professor of Biostatistics, John Hopkins Bloomberg School of Public Health
- Susan McBride, PhD, RN-BC, Nurse Informaticist and Professor, Texas Tech University Health Sciences Center
- Marc Overhage, MD, PhD, Chief Medical Informatics Officer, Siemens Health Services
- Monica Peek, MD, MPH, Assistant Professor of Medicine and Associate Director, Chicago Center for Diabetes Translation Research
- Ileana Pina, MD, MPH, Professor of Medicine and Epidemiology, Albert Einstein College of Medicine; Attending Physician, Montefiore Medical Center
- Jeremiah Schuur, MD, MHS, Attending Physician and Chief of the Division of Health Policy Translation, Brigham and Women's Hospital
- Kent Sepkowitz, MD, Deputy Physician-in-Chief for Quality and Safety, MSKCC
- Donna Slosburg, RN, BSN, Executive Director of ASC Quality Collaboration

Patient and Family Advisory Board:
This panel provided feedback on the measure concept from the patient and family perspective.
- Darlene Barkman - Children's Hospital of Philadelphia
- Ann Cannarozzo - Rochester Regional Health System
- Maureen Corcoran - Cystic Fibrosis Foundation
- Ilene Corina - PULSE (Persons United Limiting Substandards and Errors in Healthcare) of NY
- John Harris - Johns Hopkins Hospital
- Toby Levin - Suburban Hospital Patient and Family Advisory Council
- Christopher Mason - Peace Health Patient Advisory Council
- Teresa Masters - Patient and Family Centered Council, University of California, San Diego
- Lisa McDermott - National Brain Tumor Society
- Kelly Parent - Patient and Family Centered Care Program, University of Michigan Health System

- Lee Tomlinson - Center for More Compassionate Care
- Karel Shapiro - Rochester General Hospital

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?** Specifications for this eCQM will be reviewed and updated annually.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2018

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.

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have 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:** Not applicable