NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3541

Corresponding Measures:

De.2. Measure Title: Annual Monitoring for Persons on Long-Term Opioid Therapy (AMO)

Co.1.1. Measure Steward: Pharmacy Quality Alliance

De.3. Brief Description of Measure: The percentage of individuals 18 years of age and older who are on long-term opioid therapy and have not received a drug test at least once during the measurement year.

1b.1. Developer Rationale: The measure addresses the White House priority of combating the national opioid crisis.[1-3]Evidence suggests that the majority (~60%) of opioid abuse fatalities originate from opioids prescribed within practice guidelines.[4] This underscores the need for more information in the care and management of individuals on long-term opioid therapy to prevent or reduce occurrences of opioid-related adverse drug events (ADEs). Drug testing can assist prescribing clinicians in understanding risk levels for opioid-related ADEs and how to manage risks as part of a larger program of risk evaluation and minimization.[5] Clinicians should not dismiss patients from care based on the results from drug test. Such practice could have adverse consequences for patient safety.[6] Although unique insurance product types may reimburse drug tests differently, and drug testing requirements vary by state,[7] drug testing for individuals on long-term opioid therapy is nonetheless a standard of care.

No set of characteristics sufficiently identifies those at risk of opioid misuse or abuse, indicating the need to bolster prescribing clinicians' information regarding risks associated with long-term opioid therapy. Although substance abuse history and the use of psychotropic medications have been evidenced in multiple studies as risk factors,[8-10] not all patients will demonstrate aberrant behavior when examined by their prescribing physician. For example, one systematic review detailed a study that concluded that 49% of patients with a positive drug screen had no evidence of aberrant behavior, highlighting the fact that behavioral screening alone could potentially miss a substantial number of patients at risk for adverse opioid outcomes.[11] In another study of 248 patients who presented to the emergency department (ED) with complaints of pain, 32% admitted to some drug use but still tested positive for other unclaimed drugs; of patients who denied all drug use, 30% tested positive for drugs.[12]

The results of drug tests can inform the presence of illicit drug use and other medications. By measuring the proportion of individuals on long-term opioid therapy who have not received drug testing at any point in the measurement year, plans can identify patients that are not receiving care aligned with clinical practice guidelines.[4,6,13-15]

Citations:

1. Trump DJ. Establishing the President´s Commission on Combating Drug Addiction and the Opioid Crisis. Office of the Federal Register. 2017;82(62):16283-16285.

2. Christie C, Baker C, Cooper R, Kennedy PJ, Madras B, Bondi P. The President´s Commission on Combating Drug Addiction and the Opioid Crisis. Washington, DC: Office of the Press Secretary; 2017. https://www.whitehouse.gov/sites/whitehouse.gov/files/images/Final_Report_Draft_11-1-2017.pdf. Accessed May 2, 2018.

3. Hargan ED. Determination that a public health emergency exists. Washington, DC: Office of the Press Secretary; 2017. https://www.hhs.gov/sites/default/files/opioid%20PHE%20Declaration-no-sig.pdf. Accessed May 8, 2018.

4. Manchikanti L, Abdi S, Atluri S, et al. American Society of Interventional Pain Physicians (ASIPP) guidelines for responsible opioid prescribing in chronic non-cancer pain: Part I--evidence assessment. Pain Physician. 2012;15(3 Suppl):S1-65.

5. Heit HA, Gourlay DL. Using urine drug testing to support healthy boundaries in clinical care. J Opioid Manag. 2015;11(1):7-12. doi: 10.5055/jom.2015.0247.

6. Dowell D, Haegerich T, Chou R. CDC guideline for prescribing opioids for chronic pain — United States, 2016. Atlanta, GA: Centers for Disease Control and Prevention; 2016. http://www.cdc.gov/mmwr/volumes/65/rr/rr6501e1.htm. Accessed May 6, 2018.

7. Office for State Tribal Local and Territorial Support, Centers for Disease Control and Prevention. Menu of Pain Management Clinic Regulation. Atlanta, GA: US Department of Health and Human Services; nd. https://www.cdc.gov/phlp/docs/menu-pmcr.pdf. Accessed December 13, 2018.

8. Turner BJ, Liang Y. Drug overdose in a retrospective cohort with non-cancer pain treated with opioids, antidepressants, and/or sedative-hypnotics: interactions with mental health disorders. J Gen Intern Med. 2015;30(8):1081-1096. doi: 10.1007/s11606-015-3199-4.

9. Boscarino JA, Rukstalis M, Hoffman SN, et al. Risk factors for drug dependence among out-patients on opioid therapy in a large US health-care system. Addiction. 2010;105(10):1776-1782. doi: 10.1111/j.1360-0443.2010.03052.x.

10. Edlund MJ, Martin BC, Devries A, Fan MY, Braden JB, Sullivan MD. Trends in use of opioids for chronic noncancer pain among individuals with mental health and substance use disorders: the TROUP study. Clin J Pain. 2010;26(1):1-8. doi: 10.1097/AJP.0b013e3181b99f35.

11. Starrels JL, Becker WC, Alford DP, Kapoor A, Williams AR, Turner BJ. Systematic review: treatment agreements and urine drug testing to reduce opioid misuse in patients with chronic pain. Ann Intern Med. 2010;152(11):712-720. doi: 10.7326/0003-4819-152-11-201006010-00004.

12. Schuckman H, Hazelett S, Powell C, Steer S. A validation of self-reported substance use with biochemical testing among patients presenting to the emergency department seeking treatment for backache, headache, and toothache. Subst Use Misuse. 2008;43(5):589-595. doi: 10.1081/JA-200030572.

13. Langman L, Jannetto P. Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients. The American Association for Clinical Chemistry Academy; 2017. https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients. Accessed May 3, 2018.

14. US Department of Veterans Affairs, US Department of Defense. Management of Opioid Therapy for Chronic Pain. VA/DoD clinical practice guideline for opioid therapy for chronic pain. Washington, DC: US Department of Veterans Affairs and US Department of Defense; 2017. https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPG022717.pdf. Accessed May 11, 2018.

15. Chou R, Fanciullo GJ, Fine PG, et al. Clinical guidelines for the use of chronic opioid therapy in chronic noncancer pain. J Pain. 2009;10(2):113-130. doi: 10.1016/j.jpain.2008.10.008.

S.4. Numerator Statement: Individuals in the denominator population who have not received a drug test during the measurement year.

S.6. Denominator Statement: The target population for this measure is individuals 18 years of age and older and prescribed long-term opioid therapy during the measurement year. Individuals are excluded if they have had any claims indicating a cancer diagnosis or hospice care at any time during the measurement year.

S.8. Denominator Exclusions: The measure excludes individuals with: 1) a diagnosis of cancer at any time during the measurement year; or 2) hospice care at any time during the year.

De.1. Measure Type: Process

S.17. Data Source: Claims, Enrollment Data

S.20. Level of Analysis: Health Plan

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

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? Not applicable.

Preliminary Analysis: New Measure

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 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

- This is a measure of the percentage of individuals 18 years of age and older who are on long-term opioid therapy and have not received a drug test at least once during the measurement year.
- The developer provided a logic model depicting the relationship between drug testing, evaluating an opioid treatment plan, improving patient adherence to the plan of care, repeating annual drug tests, and reducing emergency department visits, hospitalizations, and mortality.
- The developer provided a summary of the five evidence-based clinical practice guidelines that support the importance of drug testing for patients on long-term opioid therapy.
- The eight guideline statements provided are generally rated moderate-to-strong for evidence and recommendation strength (only a couple of the recommendation statements have a low/weak support). Some of the evidence is stronger for monitoring prior to initiation and for those at higher risk, but does support either annual, periodic, or random monitoring for all patients on chronic opioid

therapy.

Exception to evidence
N/A

Questions for the Committee:

- Does the evidence support the importance of a at least one drug test for all patients without cancer on chronic opioid therapy?
- Does the logic model support the link between drug testing in these patients and improved patient outcomes?

Guidance from the Evidence Algorithm

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

The highest possible rating is "High" for evidence.

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

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

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

- The developer provided current performance rates using 2015-2016 data from seven QHP products from four QHP issuers as well as performance data from 52 Medicare PDPs during this time period.
- For QHPs, the mean performance score range was 57.9% to 84.1% in 2015 and from 58.0% to 83.0% in 2016.
- For Medicare PDPs mean performance was 69.1% and 67.3% in 2015 and 2016, respectively. Scores ranged from 48.5% to 83.9% in 2015 and from 48.3% to 80.7% in 2016.
- Performance rates suggest variation in performance and an opportunity for improvement.
- The developer also provides literature to support the opportunity for improvement of drug testing rates.

Disparities

- The developer reports stratified analyses using reference groups: Female for gender, White for race/ethnicity, 45-64 years old for age group in QHP data, 65+ years old for age group in Medicare data, and Medicare only for dual-enrolled status.
- The developer reports that for sex or age, the small QHP sample sizes restricted the ability to conduct disparities analysis. For the QHP products with sufficient sample sizes, results did not suggest disparities in care for sex or age.
- The developer reports that for Medicare PDPs, females and older adults are tested significantly less often than males or younger cohorts. Non-dual-eligible beneficiaries had a relative difference of more than 20% less drug tests than dual-eligible beneficiaries in 2015 and 2016. African Americans had a relative difference of greater than or equal to 10%, which means African Americans had significantly more drug tests that the White reference group.
- The developer noted that applicable variables used to determine performance among demographic characteristics were limited since socio-demographic characteristic coding is not standardized within administrative claims and it is unclear which entity is responsible for capturing and reporting social risk data.
- The developer includes three studies that do not support disparities by gender, sex, age, and race.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Based on the disparities testing results, should the developer consider stratifying results?

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

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

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

- Yes. This measure appears to be for quality improvement and is actionable by plans and providers.
- Substantial evidency supporting need for toxicity screening while taking long term opioids (e.g. CDC, ASIPP, etc). Minimum of once a year is accepted as base. After that frequency is not as well established.
- There is evidence to support the measure as upwards of 32% of patients in some studies denied drug use but when screened were using other drugs in addition to opiates
- This is a process measure based on plan level claims data measuring in the numerator the number of individuals 18 years and older on long term opioid therapy who have not had a drug test during the measurement year and the denominator the number of individuals 18 years and older on long term opioid therapy. A logic model was presented depicting that drug testing would decrease the number of emergency department visits, hospitalizations and fatalities. There was a systematic review and grading of the body of empirical evidence. There were 5 evidence-based clinical practice guidelines all of which included a myriad of studies which included randomized trials, observational studies, retrospective data analysis and meta analyses. They all articulated that drug testing would be beneficial in improving care. These apply directly and relate to the desired outcome at least in moderate fashion

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

- Yes. There is wide variation in reporting of drug testing, with variations from 20%-60% range, showing significant variation and room for improvement. There were stratified disparities for women, low income and persons of color.
- For this measure, lower number is more compliant. Performance scores across 7 QHP the PS ranged from 57.9% to 84.1% in 2015 and 58% to 83% in 2016. NOTE: QHP products with 500 or fewer members were excluded from analysis and any group with denominators < 30 members were excluded from analysis. Although this measure is being requested for QHP QSR currently, CMS may consider it in future for Medicare, so Medicare PDPs (18M + members) also had PS: 48.5% to 83.9% in 2015 and 48.3% to 80.7% in 2016. For both QHP and PDPs, there is a performance gap. Disparities: Due to coding standards, limited demographic characteristics were tested for disparities.

Gender (F is reference), Race/Ethnicity (W is reference), age group (45 yo-64 yo is reference for QHP and 65+ for Medicare) were measured for disparities. The Agency for Healthcare Research and Quality (AHRQ) for the National Healthcare Quality and Disparities Report's method for disparity calculation was utilized. A Z-test for the difference between two proportions was significant with alpha level of < 0.05, and relative difference between proportions was greater than 10% (p-value stat sig at alpha < 0.05 . For QHP only gender and age could be tested for disparity due to small sample sizes for race breakouts. For QHP, there was no disparity in care for 2015 and 2016. For Medicrare PDPs, disparity analyses suggest females and older adults are tested significantly less often than males and younger cohorts (p value of 0.0001 for both sets). The presence of disparities in national Medicare data set supports the need for measurement since we may expect to see the same disparities in QHP population if a larger national data set was available.

- There is a gap but it is narrowing. Be good to see 2019 data. 2015 data went from 50% range in 2015 to 80%+ in 2016 so the gap is getting smaller as more people are paying attention to this issue
- Yes, performance data was provided which demonstrated a performance gap and an opportunity for improvement. Across 7 QHP products from 4 QHP issuers that had sufficient denominators to report measure rates, performance scores ranged from 57.9% to 84.1% in 2015 and 58% to 83% in 2016. Among Medicare PDP's. scores ranged from 48.5% to 83.9% in 2015 and form 48.3 % to 80.7% in 2016. As for disparities, stratified analyses using reference groups were employed. In the QHP's. there was no disparity in sex or age. In Medicare PDP's, females and older adults were tested less than males and younger adults. African Americans were tested more than Caucasians. However, the studies did not support disparities.

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.

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

Evaluators: NQF Staff

NQF Staff Review

Review Summary

Reliability:

- The developer conducted score-level reliability testing using two data samples listed below. Reliability testing was performed using a beta-binomial model (i.e. signal-to-noise).
- Data 1: The mean reliability score was calculated across 7 QHP products in 2015 and 8 QHP products in 2016.
- Data 2: Medicare PDP reliability scores were computed using the methods of minimum denominator and volume categories. 67 Medicare PDPs were included in 2015 and 63 Medicare PDPs were included in 2016. This data is meant to supplement the QHP reliability.
- A minimum reliability score of 0.7 was used to indicate sufficient signal strength to distinguish differences in performance. The results indicate that the measure is reliable at the health plan level.
    - The mean reliability for QHP products with at least 30 members in the denominator (n=7) was 0.85 (range: 0.59 to 0.99).
    - 52 of 62 Medicare PDPs had at least 100 beneficiaries in the denominator. Mean reliability of these PDPs was 0.72. The range was not provided.

Validity:

- The developer submitted face validity testing for this new NQF submission. Note that new measures do not require empirical validity testing. The developer performed score-level, face validity testing with a technical expert panel (TEP). The TEP was comprised of three representatives from large QHP issuers and nine representatives from other stakeholder groups, such as measurement industry representatives, clinical and nonclinical experts, and patient caregivers/representatives.
- The TEP was asked whether or not they agree with the following statement: "The performance scores resulting from the measure as specified, can be used to distinguish poor plan-level quality related to the process of administrating at least one drug test during the measurement year among those with long-term opioid therapy."
- 9 of 9 participating TEP members (100%) voted that the measure has face validity as specified.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- Are there any concerns about the minimum denominator size (n=100) in the Medicare PDP testing required for acceptability reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, ability to determine meaningful differences in performance)?

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

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

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

- The elements are clearly defined. However, 1) the numerator is framed as a "negative," i.e. the percent not receiving urine testing [which may make mis-interpreted]; and,2) the denominator should represent national standard for PDMP reporting: which is generally >100meq morphine equivalent AND >90 days. Lastly 3) it was unclear which data-set was preferred if there was overlap between QHP claims data and Medicare claims data. Are these exclusive sets?
- All elements of numerator are obtained from known codes with define list of codes to be included. For denominator it is also claims based (rx claims of opioid therapy with 90 days of cumulative days' supply). The only question/comment i would like to add for denominator is that measure has 90 day supply defined as 'long-term opioid therapy is defined as at least 90 days of cumulative days' supply of any combination of opioid medications indicated for pain during measurement period... medications prescribed as pat of MAT for opioid use disorder are excluded." 2 questions on this: 1. if someone were to get 3 different opioid prescribed for 30 day supply each on Day 1 or within 1 week, then does this qualify as 90 day supply total (e.g. there is no mention of removing overlapping day supply count from 90 day supply calculation), Question 2: is there a list of drugs that could be used for pain and MAT that will be used as an exclusion (e.g .methadone, suboxone)?
- The measure seems reliable and the data elements are specific
- There are no concerns that this measure cannot be consistently implemented. Score level reliabilty testing was conducted using 2 data set samples.Testing used a beta-binomial model and mean reliability scores were calculated across 7 QHP products in 2015 and 8 QHP products in 2016. Medcare scores were computed using methods of minimum denominator and volume categories of 67 PDP's in 2015 and 63 in 2016. A minimum reliability score of 0.7 was used to indicate sufficient signal strength.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- Reliability was moderate and passed at 0.85 for QHP and 0.72 for Medicare. Further explanation of the trade off between data specificity versus reliability in medicare data-set would be useful. Most importantly, to make sure this measure does not discriminate against medicare or public plans. Also, if this reliability problem compromises the validity of any QHP vs Medicare comparison.
- Reliability estimates were computed by using methods of minimum denominator and volume categories. Reliability score of 0 implies that all variation is completely attributable to measurement error and score of 1.0 implies that all variation is caused by a real difference in performance across products. A minimum reliability score of 0.7 was used to indicate sufficient signal strength to distinguish differences in performance. For QHP products, the reliability scores ranged from 0.59 to 0.99 with mean reliability score of 0.85. Since 0.85 > 0.7, this measure can reliably distinguish performance between QHP products. For Medicare PDPs, mean reliability score was 0.72, therefore reliable to distinguish between PDP plans.
- No
- No

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

- Face validity was adequate. Question: unclear if differences in QHP and Medicare datasets impact validity for purpose of comparison? or will these groups be reported differently?

- Face validity with technical expert panel consisting of 3 reps from large QHP issuers, and 9 reps from other stakeholder groups (e.g. measurement industry reps, clinical and non-clinical experts, patient/caregiver reps). TEP members were asked if 'the performance scores resulting from the measure as specified can be used to distinguish good from poor plan-level quality.' 9 out of 9 TEP members (100%) participated in voting and agreed that the measure was valid. Due to the time the measure was being tested, there was a validity of ICD-9-CM to ICD-10-CM conversion. The SMEs reviewed the codes and agreed the converted ICD-10-CM codes were consistent with the intent of measure.
- No
- No

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

- There was no reported missing data.
- Sensitivity analysis found the effect of the HCPCS code change was negligible. Potential false positives due to generic codes that do not specify what drug is tested had little effect on measure performance (high P-values for the various tests). One data that I did not see tested for threat to validity is regards to drugs included in the measure for denominator. If drugs that could be used for detox and pain are included, is the measure valid? I don't see this tested.
- I did not see any threats to validity
- Threats to validity were discussed focusing on the specification of drug tests for the numerator. Prior to 2016, HCPCS, CPT and LOINC all allowed for identification of testing for specific drug classes. IN 2016, HCPCS consolidated 28 drug class codes into 4 generis codes. Sensitivity analysis was run and the Subject Matter Panel reviewed the analysis and found sufficient similarities that there was no threat to the validity. The performance scores were reviewed by the Technical Expert Panel as part of the face validity and 100% were in agreement that the scores can be used to distinguish poor plan-level quality related to the process of administering at least one drug test during the measurement year among individuals on long term opioid therapy.

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

- There was no global risk adjustment or stratification. However, since there were notable stratification gaps noted in 1.b, I would hope data could be stratified at the QHP or provider level.
- measure was tested for sub-group exclusion criteria impact on measure results. cancer individually, hospice individually, and cancer with hospice exclusions had minimum impact on measure (if the excluded populations were to be included then only a 1 to 4% impact on the measure). However,

cancer and palliative care will be excluded from measure. There was no risk adjustment or stratification performed.

- The exclusions are appropriate
- Risk adjustment was not included as this is a process measure. Individuals receiving opioid for MAT were excluded.

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 is extracted from administrative claims which are routinely collected for billing purposes.
- All data elements are in defined fields in electronic claims.

Questions for the Committee:

- Is the measure feasible to be specified and calculated using administrative claims data?
- Are there any concerns about missing data?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

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

- Since this is based mostly on claims data, it seems feasible--and a possible e-measure.
- All data can be collected electronically. Only concern of more clearly defining calculation of 90-day supply of opioids and how to handle opioids that could be used for pain or detox. In prescription claims, there are no diagnosis codes to confirm if opioid is being used for pain or detox.
- All data elements should be available from the lab and/or the EHR
- No concerns

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.

Planned use in an accountability program? ☒ Yes ☐ No

Accountability program details

- CMS anticipates proposing to add the measure to the Quality Rating System (QRS) beginning in 2021 as part of the 2020 QRS Call Letter process. Data collection would begin in the 2021 ratings year and CMS scoring in 2022.

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

- Feedback was received from the TEP (which included health plans – those being measured), three subject matter experts, and the CDC. The developer references the positive validity ratings of the TEP.
- Specific feedback considered during development includes: support for including all individuals on long-term opioid therapy rather than new, long term-therapy, support for inclusion of all opioids indicated for pain and exclusion of IV or epidural opioids, support for any drug test in the measurement year to count in the numerator, recommendation on which drug tests to specify in the numerator, and recommendation to use the term "long-term opioid therapy".

Additional Feedback: N/A

Questions for the Committee:

- Can the performance results be used to further the goal of high-quality, efficient healthcare?
- Has the measure been appropriately vetted by health plans?

Preliminary rating for Use: ☒ Pass ☐ No Pass

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 and has not been implemented. The developer references studies to support that the measure is actionable by providers and plans.

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

There are no unexpected findings from implementation to report, as this is a new measure.

Potential harms

There are no anticipated harms from using this measure.

Additional Feedback:

N/A

Questions for the Committee:

- Can health plans use the measure results to improve care?
- Any concern that drug screening will have a negative impact on care?

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

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

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

- I would prefer that this measure be intended for quality improvement rather than public reporting. As a public report measure, there are problems with population mix, funding mix and stratification, which makes for apples-to-apples interpretation problematic. As well as marginal reliability in medicare population. Also, this measure should report %tested rather than %not-tested--this is just my peeve.
- It will be used for 2020 QHP QRS and CMS may use it in future for Medicare population. It will be a transparent measure that I believe members will be able to view to see which QHP plan is better or worse
- This is reported at the plan level so would be available publicly
- CMS anticipates proposing to add this measure to the Quality Rating System in 2021 so it is not being publicly reported. Feedback was provided in 3 different methods. Those measured were given performance results, were given an opportunity to provide feedback and this feedback was considered when changes were incorporated into the measure. The TEP gave feedback which included support for including for all individuals on long term opioid therapy the inclusion of all opioids excluding those used intravenously or for epidurals.

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

- This is a good measure, but sets a higher bar than currently standard of care. It measures Utox on all persons >90 day dosing, regardless of amount. I would make like to add the floor on dosing (limited to >100meq) since the goal of this is to reduce adverse deaths (see gap)--and the

combination is a more specific. This may drive to inappropriate testing, particularly in medicare populations.

- There is data in literature supporting that physicians are likely to change opioid prescriptions based upon drug tox screening (one study with 83 cases, 69% of cases). VA has data to show 1% increase in drug tox screen can decrease patient risk of opioid related suicide or overdose event by 1%
- I do not see any harms in usability nor impediments to use
- A credible rationale was provided based on the body of evidence-based literature showing that use of drug testing can elicit illegal use of drugs which will not necessarily be discerned from observation of the individual. The only harm could be incorrect interpretation of the drug results.

Criterion 5: Related and Competing Measures

Related or competing measures

Related

- 1617 Patients Treated with an Opioid who are Given a Bowel Regimen
- 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
- 3316 Safe Use of Opioids – Concurrent Prescribing
- 3389 Concurrent Use of Opioids and Benzodiazepines (COB)

Harmonization

The above listed measures share similar populations of interest (denominator, patients receiving opioids) with the submitted measure but have different areas of focus (numerator).

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

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

- Harmonization did not seem needed or was minimal
- PQA has reviewed the other 6 other opioid related metrics, and none of them have drug tox screening as part of the measure. They have harmonized the measure to include cancer and hospice exclusions just as the 4 other ambulatory opioid related NQF metrics.
- There are a number of measures identified as competing, however, they do not address what this measure addresses i.e. other drug use while on opioids. The other measures are: 1617 Patients Treated with an Opioid who are Given a Bowel Regimen 9 • 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 • 3316 Safe Use of Opioids – Concurrent Prescribing • 3389 Concurrent Use of Opioids and Benzodiazepines (COB)
- Related but not competing measures exist.

Public and Member Comments

No comments or member support/non-support choices have been submitted as of 01/23/2020.

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3541

Measure Title: Annual Monitoring for Persons on Long Term Opioid Therapy

Type of measure:

☒ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☐ **Facility** ☒ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

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

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

   None.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - The developer conducted score-level reliability testing using a beta-binomial model (i.e. signal-to-noise).
   - The mean reliability score was calculated across 7 QHP products in 2015 and 8 QHP products in 2016.
   - Medicare PDP reliability scores were also computed using the methods of minimum denominator and volume categories. 67 Medicare PDPs were included in 2015 and 63 Medicare PDPs were included in 2016. This data is meant to supplement the QHP reliability.

7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - The mean reliability for QHP products with at least 30 members in the denominator (n=7) was 0.85 (range: 0.59 to 0.99).
   - 52 of 62 Medicare PDPs had at least 100 beneficiaries in the denominator. Mean reliability of these PDPs was 0.72. The range was not provided.

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

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

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

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

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

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

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

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

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

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

No concerns. Individuals with hospice or cancer (except non-melanoma skin cancer) are excluded. Exclusions are based on guidelines, alignment with other opioid-related measures, and expert guidance. Data shows performance rates are similar with and without exclusions.

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

    Submission document: Testing attachment, section 2b4.

    No concerns. Data provided shows significant variation in mean performance across QHP products (range 58.3% to 83%). For Medicare PDPs, 36.5% (19/50) of PDPs had rates significantly lower than the mean, and 42.3% (22/52) of PDPs had rates significantly greater than the mean.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Testing attachment, section 2b5.
    N/A

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

    Submission document: Testing attachment, section 2b6.

    The developer states missing data are not a concern since the measure relies on final paid claims. The developer also analyzed the impact of HCPCS code changes and found no threat to validity.

16. **Risk Adjustment**

    16a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

    16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

    ☐ Yes ☐ No ☒ Not applicable

    16c. **Social risk adjustment:**

    16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

    16c.2 Conceptual rationale for social risk factors included? ☐ Yes ☒ No

    16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☒ No

    16d.**Risk adjustment summary:**

    16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

    16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

    16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

    16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

    16d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

    16e**. Assess the risk-adjustment approach**

    N/A

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**

18. **Method of establishing validity of the measure score:**

☒ **Face validity**

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

☐ **N/A (score-level testing not conducted)**

19. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

- The developer performed score-level, face validity testing with a technical expert panel (TEP). The TEP was comprised of three representatives from large QHP issuers and nine representatives from other stakeholder groups, such as measurement industry representatives, clinical and nonclinical experts, and patient caregivers/representatives.
- The TEP was asked whether or not they agree with the following statement: "The performance scores resulting from the measure as specified, can be used to distinguish poor plan-level quality related to the process of administrating at least one drug test during the measurement year among those with long-term opioid therapy."

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

Submission document: Testing attachment, section 2b2.3

- 9 of 9 participating TEP members (100%) voted that the measure has face validity as specified.

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

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

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**

NOTE that data element validation from the literature is acceptable.

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

☐ **Yes**

☐ **No**

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

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

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

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

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

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

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

ADDITIONAL RECOMMENDATIONS

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

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

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

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

AMO_evidence_attachment_2019-10-21_FV-637076083541421365.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

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

Measure Title: Annual Monitoring for Persons on Long-Term Opioid Therapy (AMO)

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

Date of Submission: 10/28/2019

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

☐ 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: The percentage of individuals 18 years of age and older who are on long-term opioid therapy and have not received a drug test at least once during the measurement year.

☐ 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.

Patient is evaluated and determined to be a candidate for long-term opioid therapy → **Clinician administers a drug test** → Clinician includes information from drug test to evaluate opioid treatment plan, and if needed refers patient for opioid use disorder → Patient improves adherence to the plan of care → Repeat drug tests

at least annually and reevaluate the opioid treatment plan as necessary → Reduction in emergency department visits, hospitalizations, and mortality associated with opioid adverse drug events (ADEs).

The measure *Annual Monitoring for Persons on Long-Term Opioid Therapy (AMO)* focuses on individuals on long-term opioid therapy who have not received a drug test at least once in the measurement year. The results of drug tests are essential sources of information for providers of patients receiving long-term opioid therapy. Routine drug screenings can inform providers of aberrant drug-related behaviors, which can then influence referrals for substance use disorder, and can inform providers when more patient education is warranted to prevent potential drug-drug interactions if undisclosed drugs are discovered through test results. Additionally, drug tests can inform the provider if prescribed opioids are not evidenced by test results, indicating potential diversion or the need to change the treatment regimen to optimize patient outcomes when opioids are needed but not used. Evidence suggests that the majority (~60%) of opioid abuse fatalities originate from opioids prescribed within practice guidelines.[1] This underscores the need for more information on the care and management of patients on long-term opioid therapy to prevent or reduce occurrences of opioid-related ADEs. The importance of drug testing for patients on long-term opioid therapy is supported by five evidence-based clinical practice guidelines that recommend drug testing at the initiation of long-term opioid therapy and periodically after that. These guidelines and recommendations are detailed below in 1a.3.

Aberrant Medication-Taking or Substance Use Behaviors

Data indicate that 2 million people in the US have an opioid use disorder.[2] Because substance abuse is a risk factor for opioid-related ADEs, administering drug tests should improve patient safety by providing more information to prescribing clinicians. Additionally, patients most vulnerable to addiction, abuse, and overdoses are more likely to receive prescriptions for opioids.[3] For example, a systematic literature review found that among patients receiving opioid treatment for chronic back pain, there was a 36% to 56% prevalence rate of lifetime substance use disorders, a current substance abuse rate of 43%, and a 5% to 24% rate of aberrant medication-taking behaviors.[4]

Administering drug tests can provide health plans and clinicians with information indicative of opioid misuse. For example, in a study of 248 patients presenting to an emergency department seeking narcotics treatment for pain, urine drug tests revealed that 32% of patients admitted to some drug use but still tested positive for other unclaimed drugs; of patients who denied all drug use, 30% tested positive for drugs.[5] In another study among 120 chronic pain patients who received urine drug testing (UDT) at an urban pain management clinic, 54% were nonadherent to their prescribed opiate regimen.[6] Nonadherence was defined as absence or inappropriate level of prescribed controlled medication, presence of additional non-prescribed controlled substances, presence of illicit substances, or presence of adulterant in the urine sample. Of these who were nonadherent, 23% tested negative for one or more of their prescribed controlled medications.[6] In a larger sample of 2,741 patients on long-term opioid therapy who had UDT, 10.7% tested negative for opioids.[7] In a sample of 5,420 UDTs from 3,809 patients, 30% of UDTs showed aberrant results, including prescribed opioid non-detection (12.3%).[8] Reasons for absence of opioids could include not taking the medication in the previous several days, diluted or adulterated urine, altered opioid metabolism, or drug diversion.[7]

Polysubstance Use

Opioids and benzodiazepines can depress respiration frequency, especially when used in combination.[9] The significant concurrent use of these two drugs,[5,9] and the associated risk, is evidenced by the fact that the combination of pharmaceutical opioids and benzodiazepines is a common cause of polysubstance overdose deaths.[10] Between 15% and 25% of patients on long-term opioid therapy filled at least one benzodiazepine prescription, and 7% to 13% received greater than a 90-day supply.[11] Rates of concurrent opioid and benzodiazepine use are estimated at 45% among patients on long-term opioid therapy with comorbid depression.[11] Administering drug tests inform prescribing clinicians of risk factors for ADEs associated with opioid therapy, such as polysubstance use.

Evidence of improving health outcomes

More frequent visits in which drug testing is utilized, and retesting of patients with aberrant results, can improve adherence to prescription opioids. For example, retesting patients with aberrant results was shown to improve compliance in 63.6% (49/77) of patients in one study,[12] and 45% (77/171) of patients in another study.[13] After 14 office visits in which drug screening was conducted, there was a decreased trend in the percentage of patients who tested positive for one or more illicit drugs (23% vs. 9%; p <.0001).[14] Monitoring with five or more visits improved adherence to prescribed medications by 1-8%.[15]

One study used data from the US Department of Veterans Affairs (VA) to examine drug testing implementation following the release of the 2010 VA/DoD Clinical Practice Guidelines for the Management of Opioid Therapy for Chronic Pain. It found that independent of patient-level risk factors, higher average levels of administered drug tests within VA facilities predicted a significant reduction in risk of prescription opioid-related suicide, and overdose events. This study further suggests that for every additional 1% of opioid-prescribed patients that receive drug test monitoring, there is up to a 1% reduction in patient risk of an opioid-related suicide or overdose event.[16]

Actionable by plans/providers

Results from the process of drug screening patients on long-term opioid therapy can influence providers' plans of care. In one study, 40% of patients were referred for evaluation and treatment by behavioral health and addiction specialists due to aberrant drug testing results.[17] Among patients with nonadherent drug test results, treatment changes were planned in 69% of patients.[18] In another study, patients who did not resolve aberrant behavior after a year of structured monitoring with drug testing were either transitioned into additional treatment (22/94 patients) or discharged from care (72/94 patients).[13]

Literature suggests that multi-component education and resource interventions for physicians can improve guideline-concordant care. In one study, at one-year post-intervention, patients with an intervention physician were more likely than those with control physicians to receive guideline-concordant care (65.9% versus 37.8%; p <.001).[19] Specific to UDT, patients were more likely to have received at least one drug test within one year if they had an intervention physician, compared to a control physician (74.6% versus 57.9%; p <.001). Among a federally qualified health center and primary care clinics, implementation of a chronic pain protocol and the development of electronic reports to track provider adherence to the protocol led to an 18.3% increase in the number of patients who had a urine drug screen over 12 months.[20] In community practices and Veterans Administration medical settings, 46%-50% of physicians reported drug testing patients on opioids (both new and maintained) more frequently at follow-up after receiving provider education; the observed mean frequency of drug tests increased significantly for patients being maintained on opioids (mean = 3.8-4.1; p = 0.04).[21]

Clinical practice guidelines aid in provider decision-making and can improve rates of drug testing in patients on long-term opioid therapy. In the VA, there was an increase in drug testing from 2010 to 2013 (29% versus 42%) following the first VA opioid therapy guideline recommendations.[16] In another study, results showed that rates of drug testing increased in primary care clinics (9.2% to 17.3%) following the dissemination of guidelines.[22]

Citations:

1. Manchikanti L, Abdi S, Atluri S, et al. American Society of Interventional Pain Physicians (ASIPP) guidelines for responsible opioid prescribing in chronic non-cancer pain: Part I--evidence assessment. *Pain Physician.* 2012;15(3 Suppl):S1-65.
2. Center for Behavioral Health Statistics and Quality. *Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19-5068, NSDUH Series H-54).* Rockville, MD 2019. https://store.samhsa.gov/system/files/nsduhffr2018.pdf. Accessed September 23, 2019.
3. Hackman DT, Greene MS, Fernandes TJ, Brown AM, Wright ER, Chambers RA. Prescription drug monitoring program inquiry in psychiatric assessment: detection of high rates of opioid prescribing to a dual diagnosis population. *Journal of Clinical Psychiatry.* 2014;75(7):750-756. doi: 10.4088/JCP.14m09020.

4. Martell BA, O'Connor PG, Kerns RD, et al. Systematic review: opioid treatment for chronic back pain: prevalence, efficacy, and association with addiction. *Annals of Internal Medicine.* 2007;146(2):116-127.
5. Schuckman H, Hazelett S, Powell C, Steer S. A validation of self-reported substance use with biochemical testing among patients presenting to the emergency department seeking treatment for backache, headache, and toothache. *Substance Use and Misuse.* 2008;43(5):589-595. doi: 10.1081/JA-200030572.
6. Matteliano D, Chang YP. Describing prescription opioid adherence among individuals with chronic pain using urine drug testing. *Pain Management Nursing.* 2015;16(1):51-59. doi: 10.1016/j.pmn.2014.04.001.
7. Turner JA, Saunders K, Shortreed SM, et al. Chronic opioid therapy risk reduction initiative: impact on urine drug testing rates and results. *Journal of General Internal Medicine.* 2014;29(2):305-311. doi: 10.1007/s11606-013-2651-6.
8. Turner JA, Saunders K, Shortreed SM, et al. Chronic opioid therapy urine drug testing in primary care: prevalence and predictors of aberrant results. *Journal of General Internal Medicine.* 2014;29(12):1663-1671. doi: 10.1007/s11606-014-3010-y.
9. Jones JD, Mogali S, Comer SD. Polydrug abuse: a review of opioid and benzodiazepine combination use. *Drug and Alcohol Dependence.* 2012;125(1-2):8-18. doi: 10.1016/j.drugalcdep.2012.07.004.
10. Calcaterra S, Glanz J, Binswanger IA. National trends in pharmaceutical opioid related overdose deaths compared to other substance related overdose deaths: 1999-2009. *Drug and Alcohol Dependence.* 2013;131(3):263-270. doi: 10.1016/j.drugalcdep.2012.11.018.
11. Turner BJ, Liang Y. Drug overdose in a retrospective cohort with non-cancer pain treated with opioids, antidepressants, and/or sedative-hypnotics: interactions with mental health disorders. *Journal of General Internal Medicine.* 2015;30(8):1081-1096. doi: 10.1007/s11606-015-3199-4.
12. Knezevic NN, Khan OM, Beiranvand A, Candido KD. Repeated quantitative urine toxicology analysis may improve chronic pain patient compliance with opioid therapy. *Pain Physician.* 2017;20(2S):S135-S145.
13. Barth KS, Becker WC, Wiedemer NL, et al. Association between urine drug test results and treatment outcome in high-risk chronic pain patients on opioids. *Journal of Addiction Medicine.* 2010;4(3):167-173. doi: 10.1097/ADM.0b013e3181c379ed.
14. Pesce A, West C, Rosenthal M, et al. Illicit drug use in the pain patient population decreases with continued drug testing. *Pain Physician.* 2011;14(2):189-193.
15. Yee DA, Hughes MM, Guo AY, et al. Observation of improved adherence with frequent urine drug testing in patients with pain. *J Opioid Manag.* 2014;10(2):111-118. doi: 10.5055/jom.2014.0200.
16. Brennan PL, Del Re AC, Henderson PT, Trafton JA. Healthcare system-wide implementation of opioid-safety guideline recommendations: The case of urine drug screening and opioid-patient suicide- and overdose-related events in the Veterans Health Administration. *Translational Behavioral Medicine.* 2016;6(4):605-612. doi: 10.1007/s13142-016-0423-7.
17. Gilbert JW, Wheeler GR, Mick GE, et al. Urine drug testing in the treatment of chronic noncancer pain in a Kentucky private neuroscience practice: the potential effect of Medicare benefit changes in Kentucky. *Pain Physician.* 2010;13(2):187-194.
18. Morasco BJ, Krebs EE, Cavanagh R, Hyde S, Crain A, Dobscha SK. Treatment changes following aberrant urine drug test results for patients prescribed chronic opioid therapy. *J Opioid Manag.* 2015;11(1):45-51. doi: 10.5055/jom.2015.0251.
19. Liebschutz JM, Xuan Z, Shanahan CW, et al. Improving Adherence to Long-term Opioid Therapy Guidelines to Reduce Opioid Misuse in Primary Care: A Cluster-Randomized Clinical Trial. *JAMA Intern Med.* 2017;177(9):1265-1272. doi: 10.1001/jamainternmed.2017.2468.
20. Downes JM, Klepser DG, Foster J, Nelson M. Development of a standardized approach for managing opioids in adults with chronic noncancer pain. *American Journal of Health-System Pharmacy.* 2018;75(5):321-326. doi: 10.2146/ajhp161012.
21. Larson MJ, Browne C, Nikitin RV, et al. Physicians report adopting safer opioid prescribing behaviors after academic detailing intervention. *Substance Abuse.* 2018:1-7. doi: 10.1080/08897077.2018.1449175.
22. Chen JH, Hom J, Richman I, Asch SM, Podchiyska T, Johansen NA. Effect of opioid prescribing guidelines in primary care. *Medicine (Baltimore).* 2016;95(35):e4760. doi: 10.1097/MD.0000000000004760.

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.)

****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.

Per NQF direction above, responses are provided for only one section of either 1a.2, 1a.3 or 1a.4. We responded to 1a.3.

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)

x Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

x 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**	• **Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients** • **Langman L & Jannetto P** • **2017** • **Langman L, Jannetto P. Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients. The American Association for Clinical Chemistry Academy; 2017. https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients. Accessed August 27, 2018. (page 7)** • **https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients.**
Quote the guideline or recommendation verbatim about the process, structure or	Evidence-Based Recommendation #1: "Testing biological specimens for drugs/drug metabolites is recommended and effective for detecting the use of relevant over-the-counter,

intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	prescribed and non-prescribed drugs, and illicit substances in pain management patients. Laboratory testing does not specifically identify most other outcomes, but should be used in conjunction with additional information to detect other outcomes in pain management patients." Consensus-Based Expert Opinion #1: "Based on level II evidence, baseline drug testing should be performed prior to initiation of acute or chronic controlled substance therapy. In addition, random drug testing should be performed at a minimum of one to two times a year for low-risk patients (based on history of past substance abuse/addiction, aberrant behaviors, and opioid risk screening criteria), with increasing frequency for higher-risk patients prescribed controlled substances."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Evidence-Based Recommendation #1: Evidence = I: Evidence includes consistent results from well-designed, well-conducted studies in representative populations. Consensus-Based Expert Opinion #1: Evidence = II: Evidence is sufficient to determine effects, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies; generalizability to routine practice; or indirect nature of the evidence.
Provide all other grades and definitions from the evidence grading system	Evidence = III: Evidence is insufficient to assess the effects on health outcomes because of the limited number of power studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information.
Grade assigned to the **recommendation** with definition of the grade	Evidence-Based Recommendation #1: Recommendation = A: The National Academy of Clinical Biochemistry (NACB) strongly recommends adoption; there is good evidence that it improves important health outcomes, and it concludes that benefits substantially outweigh harms. Consensus-Based Expert Opinion #1: Recommendation = A: The National Academy of Clinical Biochemistry (NACB) strongly recommends adoption; there is good evidence that it improves important health outcomes, and it concludes that benefits substantially outweigh harms. Note: The American Association for Clinical Chemistry (AACC) Academy is formerly the NACB. While the authors still use the NACB when referring to the grading of the recommendations, the guidelines are referred to as the AACC Academy Guidelines.
Provide all other grades and definitions from the recommendation grading system	Recommendation = B: The NACB recommends adoption; there is at least fair evidence that it improves important health outcomes, and it concludes that benefits outweigh harms. Recommendation = C: The NACB recommends against adoption; there is evidence that it is ineffective or that harms outweigh benefits.

	Recommendation = I: The NACB concludes that the evidence is insufficient to make recommendations; evidence that it is effective is lacking, of poor quality, or conflicting, and the balance of benefits and harms can't be determined.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity – 9 studies Quality – randomized trials; observational studies; retrospective data analyses; review article; systematic review.
Estimates of benefit and consistency across studies	"The evidence for specific schedules of drug testing in general is weak, mainly due to the lack of randomized clinical trials comparing the effectiveness of testing schedules or methods specifically in the chronic pain population. Existing practice guidelines make recommendations based on observational studies or expert consensus opinion."
What harms were identified?	Challenges with physician understanding what is tested for and the interpretability of the drug tests were identified. Limitations regarding the sensitivity and specificity of immunoassay tests could lead to negative impacts on patient care if a test comes back as a false negative or false positive; however, these limitations can be mitigated by implementing follow-up confirmatory testing with a mass-spectrometry-based assay.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The additional studies published since the literature review for the AACC Academy Guidelines do not change the conclusions from the SR. For additional citations, please see Chou et. al. (2009) below.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Responsible, Safe, and Effective Prescription of Opioids for Chronic Non-Cancer Pain: American Society of Interventional Pain Physicians (ASIPP) Guidelines** • **Manchikanti L, Kaye AM, Knezevic NN, et al.** • **02/02/2017** • **Manchikanti L, Kaye AM, Knezevic NN, et al. Responsible, Safe, and Effective Prescription of Opioids for Chronic Non-Cancer Pain: American Society of Interventional Pain Physicians (ASIPP) Guidelines. Pain Physician. 2017;20(2S):S3-S92. (1. page S61; 2. page S62)** • **http://www.painphysicianjournal.com/current/pdf?article=NDIwMg%3D%3D&journal=103**
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.	1. "Urine drug testing (UDT) must be implemented at initiation of opioid therapy, along with continued adherence monitoring to identify patients who are non-compliant or abusing prescription drugs or illicit drugs." 2. "In order to reduce prescription drug abuse and doctor shopping, adherence monitoring by UDT and prescription drug monitoring programs provide evidence that is essential to the identification of those patients who are non-compliant or abusing prescription drugs or illicit drugs."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	1. Evidence = Level II: Evidence obtained from at least one relevant high quality randomized controlled trial or multiple relevant moderate or low quality randomized controlled trials OR Evidence obtained from at least 2 high quality relevant observational studies or large case series for assessment of preventive measures, adverse consequences, and effectiveness of other measures. 2. Evidence = Level I-II: Level I – Strong; Evidence obtained from multiple relevant high quality randomized controlled trials for effectiveness OR Evidence obtained from multiple relevant high quality observational studies or large case series for assessment of preventive measures, adverse consequences, effectiveness of other measures. Level II – Moderate; Evidence obtained from at least one relevant high quality randomized controlled trial or multiple relevant moderate or low quality randomized controlled trials OR Evidence obtained from at least 2 high quality relevant observational studies or large case series for assessment of preventive measures, adverse consequences, and effectiveness of other measures.
Provide all other grades and definitions from the evidence grading system	Level III – Fair; Evidence obtained from at least one relevant high quality nonrandomized trial or observational study with multiple moderate or low quality observational studies OR At least one high quality relevant observational studies or large case series for assessment of preventive measures, adverse consequences, effectiveness of other measures.

	Level IV – Limited; Evidence obtained from multiple moderate or low quality relevant observational studies or Evidence obtained from moderate quality observational studies or large case series for assessment of preventive measures, adverse consequences, and effectiveness of other measures. Level V – Consensus based; Opinion or consensus of large group of clinicians and/or scientists for effectiveness as well as to assess preventive measures, adverse consequences, effectiveness of other measures.
Grade assigned to the **recommendation** with definition of the grade	1. Recommendation = Moderate: There is moderate confidence that the recommendation reflects best practice. This is based on: a) good evidence for a true net effect (e.g. benefits exceed harms); b) consistent results, with minor and/or few exceptions; c) minor and/or few concerns about study quality; and/or d) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a moderate recommendation. 2. Recommendation = Moderate to Strong: Strong – There is high confidence that the recommendation reflects best practice. This is based on: a) strong evidence for a true net effect (e.g., benefits exceed harms); b) consistent results, with no or minor exceptions; c) minor or no concerns about study quality; and/or d) the extent of the panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation. Moderate – There is moderate confidence that the recommendation reflects best practice. This is based on: a) good evidence for a true net effect (e.g. benefits exceed harms); b) consistent results, with minor and/or few exceptions; c) minor and/or few concerns about study quality; and/or d) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a moderate recommendation.
Provide all other grades and definitions from the recommendation grading system	Weak – There is some confidence that the recommendation offers the best current guidance for practice. This is based on: a) limited evidence for a true net effect (e.g., benefits exceed harms); b) consistent results, but with important exceptions; c) concerns about study quality; and/or d) the extent of panelists' agreement. Other considerations (discussed in the guideline's literature review and analyses) may also warrant a weak recommendation.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity – 14 studies Quality – Published randomized controlled trials (RCTs) which were not included in systematic reviews, meta-analyses, narrative reviews, and clinical practice guidelines concerning the use and

	safety of opioid analgesics in patients with chronic non-cancer pain.
Estimates of benefit and consistency across studies	"There are no studies showing the effectiveness of UDT for risk mitigation during opioid prescribing for pain; however, multiple studies have shown that UDT not only provides useful information about nonprescribed or illicit drugs, but also improves compliance."
What harms were identified?	"However, UDT results can be subject to misinterpretation and might sometimes be associated with practices that can harm patients with stigmatization or inappropriate termination from care."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The additional studies published since the literature review for the ASIPP Guidelines do not change the conclusions from the SR. For additional citations, please see Chou et. al. (2009) below.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **VA/DoD Clinical Practice Guideline for Opioid Therapy for Chronic Pain** • **US Department of Veterans Affairs & US Department of Defense** • **February 2017** • **US Department of Veterans Affairs, US Department of Defense. Management of Opioid Therapy for Chronic Pain. VA/DoD clinical practice guideline for opioid therapy for chronic pain. Washington, DC: US Department of Veterans Affairs and US Department of Defense; 2017. https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPG022717.pdf. Accessed August 27, 2018.** • **https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPG022717.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.	"Implementing risk mitigation strategies upon initiation of long-term opioid therapy, starting with an informed consent conversation covering the risks and benefits of opioid therapy as well as alternative therapies. These strategies and their frequency should be commensurate with risk factors and include: ongoing, random urine drug testing (including appropriate confirmatory testing); Checking state prescription drug monitoring programs; Monitoring for overdose potential and suicidality; Providing overdose education; Prescribing of naloxone rescue and accompanying education."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	"The Champions and Work Group used the GRADE system to assess the quality of the evidence base and assign a grade for the strength for each recommendation."
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	Recommendation = Strong for: The Work Group is confident that the desirable effects outweigh the undesirable effects.
Provide all other grades and definitions from the recommendation grading system	Weak for – The Work Group thinks it is likely that the desirable effects outweigh the undesirable effects, but is less confident than in a Strong recommendation. Strong against – The Work Group is confident that the desirable effects do not outweigh the undesirable effects. Weak against - The Work Group thinks it is likely that the desirable effects do not outweigh the undesirable effects, but is less confident than in a Strong recommendation.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity – 28 studies; six specific to drug testing. Quality – Systematic reviews, randomized controlled trials, cohort studies, case-control and nested case-control studies, secondary data analyses, and post-hoc pooled analyses.

Estimates of benefit and consistency across studies	"The confidence in the quality of the evidence was moderate for UDT and frequent follow-up and was low for OTAs [opioid therapy agreements]. The frequency of follow-up and monitoring should be based on patient level of risk as determined by an individual risk assessment."
What harms were identified?	"There may be some variation in patient values and preferences. Certain patients may appreciate the use of risk mitigation strategies and others may not." "More frequent follow-up of patients on chronic opioid therapy can affect access to care for all empaneled patients."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The additional studies published since the literature review for the VA/DoD Guideline do not change the conclusions from the SR. For additional citations, please see Chou et. al. (2009) below.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **CDC Guideline for Prescribing Opioids for Chronic Pain – United States, 2016** • **Dowell D, Haegerich T, Chou R** • **March 18, 2016** • **Dowell D, Haegerich T, Chou R. CDC guideline for prescribing opioids for chronic pain — United States, 2016. Atlanta, GA: Centers for Disease Control and Prevention; 2016. http://www.cdc.gov/mmwr/volumes/65/rr/rr6501e1.htm. Accessed August 27, 2018.** • **http://www.cdc.gov/mmwr/volumes/65/rr/rr6501e1.htm.**
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.	"When prescribing opioids for chronic pain, clinicians should use urine drug testing before starting opioid therapy and consider urine drug testing at least annually to assess for prescribed medications as well as other controlled prescription drugs and illicit drugs"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Evidence = Type 4: Clinical experience and observations, observational studies with important limitations, or randomized clinical trials with several major 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 3 evidence: Observational studies or randomized clinical trials with notable limitations.
Grade assigned to the **recommendation** with definition of the grade	Recommendation = Category B: Individual decision-making needed; different choices will be appropriate for different patients. Clinicians help patients arrive at a decision consistent with patient values and preferences and specific clinical situations.
Provide all other grades and definitions from the recommendation grading system	Category A: Applies to all persons; most patients should receive the recommended course of action.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity – 7 studies to update existing systematic review that included 39 studies. Quality – Randomized trials and observational studies (cohort studies, case-control studies, cross-sectional studies).
Estimates of benefit and consistency across studies	"The clinical evidence review did not find studies evaluating the effectiveness of urine drug screening for risk mitigation during opioid prescribing for pain. The contextual evidence review found that urine drug testing can provide useful information about patients assumed not to be using unreported drugs."

What harms were identified?	"Urine drug testing results can be subject to misinterpretation and might sometimes be associated with practices that might harm patients (e.g., stigmatization, inappropriate termination from care)." "Experts noted that in addition to direct costs of urine drug testing, which often are not covered fully by insurance and can be a burden for patients, clinician time is needed to interpret, confirm, and communicate results."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The additional studies published since the literature review for the CDC Guideline do not change the conclusions from the SR. For additional citations, please see Chou et. al. (2009) below.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Clinical Guidelines for the Use of Chronic Opioid Therapy in Chronic Noncancer Pain** • **Chou R, Fanciullo GJ, Fine PG, et al.** • **February 2009** • **Chou R, Fanciullo GJ, Fine PG, et al. Clinical guidelines for the use of chronic opioid therapy in chronic noncancer pain. The journal of pain : official journal of the American Pain Society. 2009;10(2):113-130. doi: 10.1016/j.jpain.2008.10.008.** • **http://www.jpain.org/article/S1526-5900(08)00831-6/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.	5.2. "In patients on COT who are at high risk or who have engaged in aberrant drug-related behaviors, clinicians should periodically obtain urine drug screens or other information to confirm adherence to the COT plan of care." 5.3. "In patients on COT not at high risk and not known to have engaged in aberrant drug-related behaviors, clinicians should consider periodically obtaining urine drug screens or other information to confirm adherence to the COT plan of care."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	5.2. Evidence = Low Quality: Evidence is insufficient to assess effects on health outcomes because of limited number or power of studies, large and unexplained inconsistency between higher quality studies, important flaws in study design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes. 5.3. Evidence = Low Quality: Evidence is insufficient to assess effects on health outcomes because of limited number or power of studies, large and unexplained inconsistency between higher quality studies, important flaws in study design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.
Provide all other grades and definitions from the evidence grading system	High Quality – Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes (at least two consistent, higher-quality randomized controlled trials, or multiple, consistent observational studies with no significant methodological flaws showing large effects). Moderate Quality – Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, size, or consistency of included studies; generalizability to routine practice; or indirect nature of the evidence on health outcomes (at least one higher-quality trial with >100 subjects; two or more higher-quality trials with some inconsistency; at least two consistent, lower-quality trials, or multiple, consistent observational studies with no significant methodological flaws showing at least moderate effects).
Grade assigned to the **recommendation** with definition of the grade	5.2. Recommendation = Strong: Based on the panel's assessment that potential benefits of following the recommendation clearly outweigh potential harms and burdens.

	5.3. Recommendation = Weak: Based on more closely balanced benefits to harms or burdens, or weaker evidence. Decisions to follow a weak recommendation could vary depending on specific clinical circumstances or patient preferences and values.
Provide all other grades and definitions from the recommendation grading system	N/A; See above.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity – 10 studies Quality – Randomized controlled trials, observational studies, and systematic reviews.
Estimates of benefit and consistency across studies	"Although there is insufficient evidence for specific recommendations about how to monitor patients on COT, there is general agreement that monitoring should routinely include assessment and documentation of pain severity and functional ability, progress towards achieving therapeutic goals, and presence of adverse effects."
What harms were identified?	"However, targeted (non-universal) urine drug screening will miss some proportion of patients who engage in aberrant drug-related behaviors, as predictors of such behaviors are relatively weak." "Interpretation of urine drug screen results is a challenge, and requires an understanding of opioid drug metabolism, pharmacokinetics and limits of laboratory testing methods."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The additional studies published since the literature review for the AACC Academy Guidelines support the recommendations made and do not change the conclusions from the SR. Additional citations: 1. Office of National Drug Control Policy. National Drug Control Strategy. 2016. https://obamawhitehouse.archives.gov/sites/default/files/ondcp/policy-and-research/2016_ndcs_final_report.pdf. Accessed March 9, 2017. 2. Trump DJ. Presidential Executive Order Establishing the President's Commission on Combating Drug Addiction and the Opioid Crisis. Washington, DC: Office of the Press Secretary; 2017. https://www.whitehouse.gov/the-press-office/2017/03/30/presidential-executive-order-establishing-presidents-commission. Accessed April 13, 2017. 3. US Department of Health and Human Services, Office of Disease Prevention and Health Promotion. National Action Plan for Adverse Drug Event Prevention. Washington, DC 2014. https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf. Accessed August 30, 2016. 4. Ducoffe AR, York A, Hu DJ, Perfetto D, Kerns RD. National Action Plan for Adverse Drug Event Prevention: recommendations for

	safer outpatient opioid use. Pain Med. 2016. doi: 10.1093/pm/pnw106. 5. Brennan PL, Del Re AC, Henderson PT, Trafton JA. Healthcare system-wide implementation of opioid-safety guideline recommendations: the case of urine drug screening and opioid-patient suicide- and overdose-related events in the Veterans Health Administration. Transl Behav Med. 2016;6(4):605-612. doi: 10.1007/s13142-016-0423-7. 6. Park T, Saitz R, Nelson K, Xuan Z, Liebschutz J, Lasser K. The association between benzodiazepine prescription and aberrant drug-related behaviors in primary care patients receiving opioids for chronic pain. Subst Abus. 2016;19(0):Ahead of print. doi: 10.1080/08897077.2016.1179242. 7. Centers for Medicare & Medicaid Services. CMS Quality Strategy 2016. Baltimore, MD: US Department of Health and Human Services; 2015. http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/CMS-Quality-Strategy.pdf. Accessed October 14, 2016. 8. Rudd RA, Aleshire N, Zibbell JE, Gladden RM. Increases in drug and opioid overdose deaths — United States, 2000–2014. Atlanta, GA: Centers for Disease Control and Prevention; 2016. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm?s_cid=mm6450a3_w. Accessed August 22, 2016.

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.

Per NQF direction above, responses are provided for only one section of either 1a.2, 1a.3 or 1a.4. We responded to 1a.3.

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

Per NQF direction above, responses are provided for only one section of either 1a.2, 1a.3 or 1a.4. We responded to 1a.3.

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

Per NQF direction above, responses are provided for only one section of either 1a.2, 1a.3 or 1a.4. We responded to 1a.3.

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

Per NQF direction above, responses are provided for only one section of either 1a.2, 1a.3 or 1a.4. We responded to 1a.3.

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.

The measure addresses the White House priority of combating the national opioid crisis.[1-3]Evidence suggests that the majority (~60%) of opioid abuse fatalities originate from opioids prescribed within practice guidelines.[4] This underscores the need for more information in the care and management of individuals on long-term opioid therapy to prevent or reduce occurrences of opioid-related adverse drug events (ADEs). Drug testing can assist prescribing clinicians in understanding risk levels for opioid-related ADEs and how to manage risks as part of a larger program of risk evaluation and minimization.[5] Clinicians should not dismiss patients from care based on the results from drug test. Such practice could have adverse consequences for patient safety.[6] Although unique insurance product types may reimburse drug tests differently, and drug testing requirements vary by state,[7] drug testing for individuals on long-term opioid therapy is nonetheless a standard of care.

No set of characteristics sufficiently identifies those at risk of opioid misuse or abuse, indicating the need to bolster prescribing clinicians' information regarding risks associated with long-term opioid therapy. Although substance abuse history and the use of psychotropic medications have been evidenced in multiple studies as risk factors,[8-10] not all patients will demonstrate aberrant behavior when examined by their prescribing physician. For example, one systematic review detailed a study that concluded that 49% of patients with a positive drug screen had no evidence of aberrant behavior, highlighting the fact that behavioral screening alone could potentially miss a substantial number of patients at risk for adverse opioid outcomes.[11] In another study of 248 patients who presented to the emergency department (ED) with complaints of pain, 32% admitted to some drug use but still tested positive for other unclaimed drugs; of patients who denied all drug use, 30% tested positive for drugs.[12]

The results of drug tests can inform the presence of illicit drug use and other medications. By measuring the proportion of individuals on long-term opioid therapy who have not received drug testing at any point in the measurement year, plans can identify patients that are not receiving care aligned with clinical practice guidelines.[4,6,13-15]

Citations:

1. Trump DJ. Establishing the President´s Commission on Combating Drug Addiction and the Opioid Crisis. Office of the Federal Register. 2017;82(62):16283-16285.

2. Christie C, Baker C, Cooper R, Kennedy PJ, Madras B, Bondi P. The President´s Commission on Combating Drug Addiction and the Opioid Crisis. Washington, DC: Office of the Press Secretary; 2017. https://www.whitehouse.gov/sites/whitehouse.gov/files/images/Final_Report_Draft_11-1-2017.pdf. Accessed May 2, 2018.

3. Hargan ED. Determination that a public health emergency exists. Washington, DC: Office of the Press Secretary; 2017. https://www.hhs.gov/sites/default/files/opioid%20PHE%20Declaration-no-sig.pdf. Accessed May 8, 2018.

4. Manchikanti L, Abdi S, Atluri S, et al. American Society of Interventional Pain Physicians (ASIPP) guidelines for responsible opioid prescribing in chronic non-cancer pain: Part I--evidence assessment. Pain Physician. 2012;15(3 Suppl):S1-65.

5. Heit HA, Gourlay DL. Using urine drug testing to support healthy boundaries in clinical care. J Opioid Manag. 2015;11(1):7-12. doi: 10.5055/jom.2015.0247.

6. Dowell D, Haegerich T, Chou R. CDC guideline for prescribing opioids for chronic pain — United States, 2016. Atlanta, GA: Centers for Disease Control and Prevention; 2016. http://www.cdc.gov/mmwr/volumes/65/rr/rr6501e1.htm. Accessed May 6, 2018.

7. Office for State Tribal Local and Territorial Support, Centers for Disease Control and Prevention. Menu of Pain Management Clinic Regulation. Atlanta, GA: US Department of Health and Human Services; nd. https://www.cdc.gov/phlp/docs/menu-pmcr.pdf. Accessed December 13, 2018.

8. Turner BJ, Liang Y. Drug overdose in a retrospective cohort with non-cancer pain treated with opioids, antidepressants, and/or sedative-hypnotics: interactions with mental health disorders. J Gen Intern Med. 2015;30(8):1081-1096. doi: 10.1007/s11606-015-3199-4.

9. Boscarino JA, Rukstalis M, Hoffman SN, et al. Risk factors for drug dependence among out-patients on opioid therapy in a large US health-care system. Addiction. 2010;105(10):1776-1782. doi: 10.1111/j.1360-0443.2010.03052.x.

10. Edlund MJ, Martin BC, Devries A, Fan MY, Braden JB, Sullivan MD. Trends in use of opioids for chronic noncancer pain among individuals with mental health and substance use disorders: the TROUP study. Clin J Pain. 2010;26(1):1-8. doi: 10.1097/AJP.0b013e3181b99f35.

11. Starrels JL, Becker WC, Alford DP, Kapoor A, Williams AR, Turner BJ. Systematic review: treatment agreements and urine drug testing to reduce opioid misuse in patients with chronic pain. Ann Intern Med. 2010;152(11):712-720. doi: 10.7326/0003-4819-152-11-201006010-00004.

12. Schuckman H, Hazelett S, Powell C, Steer S. A validation of self-reported substance use with biochemical testing among patients presenting to the emergency department seeking treatment for backache, headache, and toothache. Subst Use Misuse. 2008;43(5):589-595. doi: 10.1081/JA-200030572.

13. Langman L, Jannetto P. Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients. The American Association for Clinical Chemistry Academy; 2017. https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients. Accessed May 3, 2018.

14. US Department of Veterans Affairs, US Department of Defense. Management of Opioid Therapy for Chronic Pain. VA/DoD clinical practice guideline for opioid therapy for chronic pain. Washington, DC: US Department of Veterans Affairs and US Department of Defense; 2017. https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPG022717.pdf. Accessed May 11, 2018.

15. Chou R, Fanciullo GJ, Fine PG, et al. Clinical guidelines for the use of chronic opioid therapy in chronic noncancer pain. J Pain. 2009;10(2):113-130. doi: 10.1016/j.jpain.2008.10.008.

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis. *(<u>This is required for maintenance of endorsement</u>. 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.*

Definitions

In this submission, the term “issuer” refers to an individual insurance company or insurance organization. The term “product” refers to a package of health coverage benefits that are offered using a particular network type (i.e., health maintenance organization, preferred provider organization, exclusive provider organization, point of service, or indemnity).[1]

Analytic Processes

Performance scores on the measure as specified are below. The measured entities include QHPs, which operate in the Health Insurance Exchanges. As a condition of participation, eligible QHPs are required to collect and submit quality measure data. CMS calculates quality ratings based on the data submitted and Exchanges are required to display QHP overall quality ratings and three summary indicator ratings to assist in consumer

selection of a QHP offered on an Exchange. The measure described in this form is being considered for inclusion for this program, known as the QRS. In order for the testing to reflect how the measure would perform if it were implemented in the QRS, testing methodology aligned with the 2018 Quality Rating System Measure Technical Specifications.[2]

- QHP products with 500 or fewer total members were excluded from all analyses, and
- Denominators had to have at least 30 members in order to show the results of analyses.

The 501 member and 30 minimum denominator rules are not part of the measure specifications. The analyses followed these rules to reflect steps that would be taken if the measure were implemented into the QRS (QHP data). As noted previously, additional measure testing was conducted on Medicare PDPs to leverage the larger sample of available Medicare data, although the measure is not currently planned for implementation in the Medicare program. The 501 member and 30 minimum denominator rules were not applied to the Medicare data since the rules are specific to the QRS (QHP data).

Performance Scores (Lower scores = better performance)

Overall, across seven QHP products from four QHP Issuers that had sufficient denominators to report measure rates, the performance scores ranged from 57.9% to 84.1% in 2015, and from 58.0% to 83.0% in 2016 (see below). Among Medicare PDPs, the performance scores ranged from 48.5% to 83.9% in 2015, and from 48.3% to 80.7% in 2016 (see below). The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries on long-term opioid therapy.

RESULTS:

QHP Issuer 1, 2015-2016

The issuer data used to calculate the measure represents 289,136 members and 3 QHP products in 2015, and 223,427 members and 3 QHP products in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / A / 45 / 33 / 73.3%

2016 / A / 49 / 40 / 81.6%

2015 / B / 1,411 / 1,187 / 84.1%

2016 / B / 1,299 / 1,040 / 80.1%

2015 / C / Insufficient denominator size for calculation

2016 / C / 47 / 39 / 83.0%

QHP Issuer 2, 2015-2016

The issuer data used to calculate the measure represents one product with 45,537 members in 2015, and 30,128 members in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / A / 944 / 655 / 69.4%

2016 / A / 702 / 437 / 62.3%

QHP Issuer 3, 2015-2016

The issuer data used to calculate the measure represents two products in 2015 representing 14,093 members, and three products in 2016 representing 75,637 members. Product C did not exist in 2015 and did not have sufficient denominator size to be presented in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / A / 74 / 53 / 71.6%

2016 / A / 462 / 268 / 58.0%

2015 / B / 38 / 22 / 57.9%

2016 / B / 391 / 233 / 59.6%

2016 / C / Insufficient denominator size for calculation

QHP Issuer 4, 2015-2016

The issuer data used to calculate the measure represents one product with 3,107 members in 2015, and 2,077 members in 2016.

Year / Product / Denominator / Numerator / Rate

2015 / A / 61 / 37 / 60.7%

2016 / A / 36 / 21 / 58.3%

Medicare PDPs, 2015-2016

The Medicare data used to calculate the measure represents 18,257,146 PDP beneficiaries in the 2015 measurement period and 18,945,015 in the 2016 measurement period. Testing results here include plans with at least 100 beneficiaries in the denominator due to reliability analyses which suggested that at least 100 beneficiaries in the denominator are needed to produce results with a reliability score of at least 0.7. For more information, please see section 2a2.4 of the testing form.

Year / n / Mean / Min / P10 / P25 / P50 / P75 / P90 / Max

2015/ 52 / 69.1% / 48.5% / 57.6% / 63.7% / 70.3% / 74.4% / 78.4% / 83.9%

2016/ 52 / 67.3%/ 48.3% / 55.8% / 62.8% / 68.7% / 73.1% / 75.9% / 80.7%

Citations:

1. Centers for Medicare & Medicaid Services. Federal Definitions for Health Insurance Products and Plans. Baltimore, MD: US Department of Health and Human Services; 2016. https://www.cms.gov/CCIIO/Resources/Training-Resources/Downloads/product-vs-plan-ppt.pdf. Accessed June 12, 2018.

2. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

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.

Testing results provided in 1b.2 are consistent with the literature, which suggests room for improvement in the rates of drug testing.

Recently published data suggest ample room for improvement in the rates of drug testing in the management of individuals on long-term opioid therapy. For example, a study of 1,612 patients on long-term opioid use (>3 monthly prescriptions in six months) for noncancer chronic pain treated by eight primary care practices revealed that 8% had a drug test at any point during the four years of the study period.[1] Additionally, 24% of the highest-risk patients received drug test.[1] Rates of drug testing also vary among subgroups of the long-term opioid therapy population. For example, among a sample of veterans who received opioid prescriptions for three or more consecutive months during a one-year period, 52% received drug testing over the course of a year.[2] At a different US Department of Veterans Affairs (VA) facility, rates of drug testing over a 12 month period for patients on 90 consecutive days of opioid use were 18% among those without a substance use disorder and 47% among those with a substance use disorder.[3] Although the rates of drug testing appear higher among certain subgroups, overall performance is suboptimal, as demonstrated by these disparate statistics.

Variation in rates of drug testing for individuals on long-term opioid therapy exists between similar sites of care. A one-year retrospective review of electronic health records of noncancer pain patients receiving greater

than three monthly prescriptions in six months from three safety-net primary care clinics revealed significant variation in the rates of drug testing (P <.001). The rates of drug testing in 12 months were 31% (site 1), 35% (site 2), and 67% (site 3).[4] A recent study, using data from VA facilities, examined drug testing implementation following the release of the 2010 VA/DOD clinical practice guidelines for opioid therapy for pain. This study found that, independent of patient-level risk factors, higher average levels of administered drug tests within VA facilities predicted a significant reduction in risk of prescription opioid-related suicide and overdose events. This study further suggests that for every additional one percent of opioid-prescribed patients that receive drug test monitoring, there is up to a one percent reduction in patient risk of an opioid-related suicide or overdose event.[5] Literature suggests that multicomponent education and resource interventions for physicians can improve guideline concordant care. In one study, at one year post-intervention patients with an intervention physician were more likely than those with control physicians to receive guideline-concordant care (65.9% versus 37.8%; p <.001).[6] Specific to urine drug testing, patients were more likely to have received at least one drug test within one year if they had an intervention physician, compared to a control physician (74.6% versus 57.9%; p <.001).

Citations:

1. Starrels JL, Becker WC, Weiner MG, Li X, Heo M, Turner BJ. Low use of opioid risk reduction strategies in primary care even for high risk patients with chronic pain. J Gen Intern Med. 2011;26(9):958-964. doi: 10.1007/s11606-011-1648-2.

2. Sekhon R, Aminjavahery N, Davis CN, Jr., Roswarski MJ, Robinette C. Compliance with opioid treatment guidelines for chronic non-cancer pain (CNCP) in primary care at a Veterans Affairs Medical Center (VAMC). Pain Med. 2013;14(10):1548-1556. doi: 10.1111/pme.12164.

3. Morasco BJ, Duckart JP, Dobscha SK. Adherence to clinical guidelines for opioid therapy for chronic pain in patients with substance use disorder. J Gen Intern Med. 2011;26(9):965-971. doi: 10.1007/s11606-011-1734-5.

4. Lange A, Lasser KE, Xuan Z, et al. Variability in opioid prescription monitoring and evidence of aberrant medication taking behaviors in urban safety-net clinics. Pain. 2015;156(2):335-340. doi: 10.1097/01.j.pain.0000460314.73358.ff.

5. Brennan PL, Del Re AC, Henderson PT, Trafton JA. Healthcare system-wide implementation of opioid-safety guideline recommendations: the case of urine drug screening and opioid-patient suicide- and overdose-related events in the Veterans Health Administration. Transl Behav Med. 2016;6(4):605-612. doi: 10.1007/s13142-016-0423-7.

6. Liebschutz JM, Xuan Z, Shanahan CW, et al. Improving Adherence to Long-term Opioid Therapy Guidelines to Reduce Opioid Misuse in Primary Care: A Cluster-Randomized Clinical Trial. JAMA internal medicine. 2017;177(9):1265-1272. doi: 10.1001/jamainternmed.2017.2468.

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.*

Differences in performance among demographic characteristics were determined using the limited applicable variables included in testing data. At this time, information required to calculate certain socio-demographic characteristics (e.g., race/ethnicity) is not coded in a standard manner within administrative claims.[1] Further, there is a lack of clarity regarding which entity (e.g., physician, group, plan, and/or employer) is responsible for capturing and reporting these data.[1] Other health plan measures (e.g., HEDIS quality measures) do not currently collect or report quality performance data stratified by sociodemographic factors.[2]

In the stratified analyses for the measure, female is the reference group for gender, white is the reference group for race/ethnicity, the age group 45-64 is the reference for QHP data, 65+ is the reference group for age for Medicare, and Medicare only is the reference group for dual enrolled status. Results may be interpreted as better, worse, or the same as a reference group.

In order to assess whether disparities in measure performance exist between subpopulations of the measure cohort, we used the method employed by the Agency for Healthcare Research and Quality (AHRQ) for the National Healthcare Quality and Disparities Report. Disparities statistics were only calculated when the comparison and reference denominators both had at least 30 members in the denominator.[1] Disparities between pairs of population groups were considered identified if the following criteria were met:

1. a Z-test for the difference between two proportions, using a pooled estimate of the variance, was significant with an alpha level of less than 0.05, and

2. the relative difference between proportions was greater than 10%

P-Value = statistically significant at the alpha <0.05 level two-tailed Z-test)

Relative Difference = [(Comparison group measure score – Reference group measure score) / Reference group measure score] * 100.

Results

Performance scores on the measure as specified are below, stratified by subpopulation. In the QHP data, the small sample sizes of many of the products restricted the ability to conduct disparities analysis for sex or age. Of the QHP products that could report measure rates stratified by sex and age, results did not suggest disparities in care in either 2015 or 2016 (data not shown for non-significant results). Using data from Medicare PDPs, disparities analyses suggest that females and older adults are tested significantly less often than males and younger cohorts. The presence of disparities in the national Medicare data set supports the need for measurement since we may expect to see the same disparities in the QHP population if we were to have access to a nationally representative data set.

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

2. National Committee for Quality Assurance. HEDIS® 2017 Volume 2 Technical Specifications for Health Plans. Washington, DC: National Committee for Quality Assurance; 2017.

RESULTS:

Medicare Part D Prescription Drug Plans – 2015 & 2016: Rates by Demographics

The following display the demographic characteristics of the denominator and numerator from Medicare claims data for beneficiaries in 2015 and 2016. Analyses suggest that females and older adults are tested significantly less often than males and younger cohorts.

2015 - Dual-Eligible Status

Significant differences in measure rate were detected between non-dual-eligible and dual-eligible beneficiaries ($p < 0.001$) with non-dual-eligible beneficiaries having a relative difference of more than 20% less drug tests than dual-eligible beneficiaries, indicating a disparity in drug screening by dual-eligible status.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

Non-dual-eligible / 643,962 / 458,117 / 71.1% / Reference / Reference

Dual-eligible / 627,161 / 356,467 / 56.8%/ 20.1 / .0001

2016 - Dual-Eligible Status

Significant differences in measure rate were detected between non-dual-eligible and dual-eligible beneficiaries ($p < 0.001$) with non-dual-eligible beneficiaries having a relative difference of more than 20% less drug tests than dual-eligible beneficiaries, indicating a disparity in drug screening by dual-eligible status.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

Non-dual-eligible / 639,263 / 427,120 / 66.8% / Reference/ Reference

Dual-eligible / 596,203 / 315,512 / 52.9% / 20.8 / .0001

2015 - Sex

Significant differences in measure rates were detected between males and females ($p < 0.001$) with females having a relative difference of at least 10% less drug tests than males indicating a disparity in drug screening by sex.

Variable / Denominator / Numerator / Measure rate / Relative difference / p-value

Female / 822,883 / 548,194 / 66.6% / Reference / Reference

Male / 448,240 / 266,390 / 59.4% / 10.8/ .0001

2016 - Sex

Significant differences in measure rates were detected between males and females ($p < 0.001$) with females having a relative difference of 11% less drug tests than males indicating a disparity in drug screening by sex.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

Female / 799,443 / 499,871 / 64.5% / Reference / Reference

Male / 436,023 / 242,761 / 55.7% / 11.0% / .0001

2015 - Age

Using the 65 and older category as a reference group, there were significant differences in measure rates when compared to all other age groups ($p < 0.001$ for all 3 comparisons). In addition, the relative difference in measure rates were at least 10% higher for those 65 and older compared to all younger age groups, indicating a disparity in drug screening by age.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

18 -26 / 2,734 / 1,314 / 48.1% / 37.6 / .0001

27 - 44 / 99,685/ 40,390 / 40.5%/ 47.4 / .0001

45 - 64 / 450,605 / 219,446 / 48.7%/ 36.8 / .0001

65+ / 718,099 / 553,434 / 77.1% / Reference / Reference

2016 - Age

Using the 65 and older category as a reference group, there were significant differences in measure rates when compared to all other age groups ($p < 0.001$ for all 3 comparisons). In addition, the relative difference in measure rates were at least 10% higher for those 65 and older compared to all younger age groups, indicating a disparity in drug screening by age.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

18 -26 / 2,190 / 1,007 / 46.0%/ 36.5% / .0001

27 - 44 / 87,696 / 32,780 / 37.4% / 48.4% / .0001

45 - 64 / 427,988 / 188,857 / 44.1% / 39.1% / .0001

65+ / 717,592 / 519,988 / 72.5% / Reference / Reference

2015 - Race

Significant differences exist between all racial categories when comparing to those who identified as white ($p < 0.0001$ for all 4 comparisons). However, the relative differences in measure rates were less than 10% to indicating there is not a disparity in drug screening by race.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

White / 1,038,167/ 671,523 / 64.7% / Reference / Reference

African American / 168,779 / 98,609 / 58.4% / 9.7 /.0001

Hispanic / 25,407 / 17,135 / 67.4% / 4.3 / .0001

Other / 32,328 / 22,989 / 71.1% / 9.9 / .0001

Unknown / 6,442 / 4,328 / 67.2% / 3.9 / .0001

2016 - Race

Significant differences exist between all racial categories when comparing to those who identified as white ($p < 0.0001$ for all 4 comparisons). Except for the African American racial category, the relative differences in measure rates were not at least 10% with African-Americans having significantly more drug tests than the White racial group indicating a disparity in drug screening between the two races.

Variable / Denominator / Numerator / Measure rate/ Relative difference / p-value

White / 1,507,234 / 949,195 / 63.0% / Reference / Reference

African American / 223,598 / 123,748 / 55.3% / 12.1 / .0001

Hispanic / 33,357 / 21,254 / 63.7% / 1.2 / .0001

Other / 42,857 / 29,087 / 67.9% / 7.8 / .0001

Unknown / 11,711 / 7,683 / 65.6%/ 4.2 / .0001

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

Krishnamurthy et. al. (2016)[1] specified a logistic regression model to assess the differences between patients who received a urine drug screen in the first visit versus patients who did not receive a urine drug screen in the first visit at a university-based pain clinic. They found that predictors included diastolic blood pressure ($p = 0.0068$), calendar year of the visit ($p < 0.0001$), having a pain procedure scheduled for the second visit ($p = 0.0127$), and having an opioid prescription in the first visit ($p = 0.0031$). Similar to our results in the QHP sample, gender and age (at time of first visit) were not found to be associated with receiving a urine drug screen during the first visit. Similarly, Becker et. al. (2011)[2] did not find an association between race and receipt of a urine drug test during opioid treatment, and Oliva et. al. (2015)[3] did not find an association between sex and receipt of a urine drug screen (for either illicit drugs or non-morphine opioid compounds).

This lack of significance in sex and gender disparities could be due to small sample sizes of patients that receive drug tests, and lack of studies reporting multivariate results for the association between demographic characteristics and receiving a drug test; section 1b.3 highlights the low rates at which drug tests are administered. In the studies discussed in this section, only 8.0% of the 1,612 patients received a urine drug test in Becker et. al. (2011),[2] and just 19.8% of the 480,809 patients were screened for illicit drugs and 10.8% were screened for non-morphine opioid compounds in Olivia et. al. (2015).[3] Krishnamurthy et. al. (2016) did not report the number of individuals that received drug tests, however their total sample consisted of just 723 patients.[1]

Citations:

1. Krishnamurthy P, Ranganathan G, Williams C, Doulatram G. Impact of Urine Drug Screening on No Shows and Dropouts among Chronic Pain Patients: A Propensity-Matched Cohort Study. Pain Physician. 2016;19(2):89-100.

2. Becker WC, Starrels JL, Heo M, Li X, Weiner MG, Turner BJ. Racial differences in primary care opioid risk reduction strategies. Ann Fam Med. 2011;9(3):219-225. doi: 10.1370/afm.1242.

3. Oliva EM, Midboe AM, Lewis ET, et al. Sex differences in chronic pain management practices for patients receiving opioids from the Veterans Health Administration. Pain Med. 2015;16(1):112-118. doi: 10.1111/pme.12501.

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.)*

Not applicable.

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 not an eMeasure **Attachment:**

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:** AMO_CompleteCoding_UPDATED-637002672397479085.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).

Individuals in the denominator population who have not received a drug test during the measurement year.

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).

Individuals in the denominator who do not have at least one claim for a drug test during the measurement year will be counted in the numerator. The entire measurement year in which a member is continuously enrolled is used to calculate the measure.

A drug test is identified either through HCPCS drug test codes or through specified CPT or LOINC codes for presumptive or definitive drug screens/tests for at least one of the following targeted drug classes: amphetamines, barbiturates, benzodiazepines, cannabinoids, cocaine, and opiates/opioids.

Qualifying CPT and HCPCS drug test codes, and suggested LOINC codes, are in the attached Excel file "AMO_CompleteCoding_UPDATED" in the following sheets: "Codes-2016 Data," "Codes-2017 Data," Codes-2018 Data," and "DrugScreen_LOINC_15,16,17."

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

The target population for this measure is individuals 18 years of age and older and prescribed long-term opioid therapy during the measurement year. Individuals are excluded if they have had any claims indicating a cancer diagnosis or hospice care at any time during the measurement year.

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).

The measurement year is defined as 12 consecutive months. Continuous enrollment is defined as 11 out of 12 months enrollment in a health plan in the measurement year or enrolled with no gaps in enrollment until the month of death in the measurement year. Long-term opioid therapy is defined as at least 90 days of cumulative days' supply of any combination of opioid medications indicated for pain during the measurement period identified using prescription claims. Medications prescribed or provided as part of medication-assisted treatment for opioid use disorder are excluded from the calculation.

The target population is adults enrolled in a Qualified Health Plan (QHP) and on long-term opioid therapy.

Eligible members for this measure are those members who:

1) Are 18 years of age and older as of the first day of the measurement year.

2) Are continuously enrolled in a QHP which is defined as at least 11 out of 12 months during the measurement year or enrolled with no gaps until the date of death.

3) Have pharmacy claims indicating at least 90 days of cumulative supply of any combination of opioid medications indicated for pain during the measurement year.

Opioid medications are specified in the attached Excel file "AMO_CompleteCoding_UPDATED" in the following sheets "2016_OPIOIDFORPAINMEDICATION," "2017_OPIOIDFORPAINMEDICATION," and "2018_OPIOIDFORPAINMEDICATION."

Days' supply is calculated by summing the days' supply for every prescription during the measurement year for opioid medications indicated for pain from the above lists. Individuals qualify for the measure denominator if this sum is at least 90 days.

Note: The active ingredient of the opioid medications is limited to formulations indicated for pain and delivered through any route except intravenous (IV) or epidural (EP). These two routes are not included in this measure because they are not commonly prescribed as chronic pain medications. Medications prescribed or provided as part of medication-assisted treatment for opioid use disorder are excluded from the calculation.

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

The measure excludes individuals with: 1) a diagnosis of cancer at any time during the measurement year; or 2) hospice care at any time during the year.

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.)*

Members with a diagnosis of cancer are identified with the diagnosis codes listed below.

Cancer exclusion ICD-9 codes (for testing only):

Include 140 through 239

Omit 173.XX series

Cancer exclusion ICD-10 codes:

Include C00 through D49

Omit C44.XX series

Members with hospice care are identified with the codes listed below.

Hospice Codes 2015-2016:

Revenue Codes – 0115, 0125, 0135, 0145, 0155, 0235, 0650, 0651, 0652, 0655, 0656, 0657, 0658, 0659

CPT Codes – 99377, 99378

HCPCS Codes – G0182, G9473, G9474, G9475, G9476, G9477, G9478, G9479, Q5003, Q5004, Q50005, Q5006, Q5007, Q5008, Q5010, S9126, T2042, T043, T2044, T2045, T2046

Type of Bill (TOB) Codes – 0810, 0811, 0812, 0813, 0814, 0815, 0817, 0818, 0819, 0820, 0821, 0822, 0823, 0824, 0825, 0827, 0828, 0829, 081A, 081B, 081C, 081D, 081E, 081F, 081G, 081H, 081I, 081J, 081K, 081M, 081O, 081X, 081Y, 081Z, 082A, 082B, 082C, 082D, 082E, 082F, 082G, 082H, 082I, 082J, 082K, 082M, 082X, 082Y, 082Z

Note: A full list of codes is provided in the attached Excel file "AMO_CompleteCoding" in the sheet "Codes-2016 Data," "Codes-2017 Data," and "Codes-2018 Data."

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.

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.)*

Denominator: Individuals 18 years of age and older who are on long-term opioid therapy during the measurement year.

Create Denominator:

1. Include all individuals enrolled in a health plan for 11 of 12 months during the measurement year or enrolled with no gaps in enrollment until the month of death in the measurement year.

a. For QHPs in the Health Insurance Marketplace, switching between QHP products is considered continuous enrollment if enrollment and claims/encounter data are available for 11 of 12 months. The measure score is attributed to the last enrolled QHP product, in accordance with technical guidance specific to the Health Insurance Marketplace Quality Rating System (QRS), available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf.

2. Include individuals from step 1 who were 18 years of age or older as of the first day of the measurement year.

3. Include individuals from step 2 with a total days' supply of opioids of 90 days or more identified in pharmacy claims (section S.7).

4. Exclude individuals with any institutional or non-institutional claims indicating a cancer diagnosis during the measurement year (section S.9)

5. Exclude individuals with any institutional or non-institutional claims indicating hospice care during the measurement year (section S.9)

6. Include only unique members from step 5 in the final denominator.

Numerator: Individuals in the denominator population with no claims for drug tests during the measurement year.

Create Numerator:

7. Include individuals from the denominator who do not have any claims for a drug test during the measurement year (section S.5)

Calculate Measure Score:

8. The measure score is calculated as the number of individuals in the numerator divided by the number of individuals in the denominator multiplied by 100 (to produce a percentage).

For the Health Insurance Marketplace, members are attributed to the last QHP enrolled product during the measurement year.

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.

This measure is not based on 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.

This measure is not based on survey or patient-reported data.

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

If other, please describe in S.18.

Claims, Enrollment Data

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.

There is no data collection instrument. Individual health plans produce administrative claims in the course of providing care to health plan members.

This measure is being considered for use in the Quality Rating System (QRS) for Qualified Health Plans (QHPs). QHPs operate in the Health Insurance Exchanges, established by the Patient Protection and Affordable Care Act. As a condition of participation, eligible QHPs are required to collect and submit quality measure data. CMS calculates quality ratings based on the data submitted, and Exchanges are required to display QHP overall quality ratings and three summary indicator ratings to assist in consumer selection of a QHP offered on an Exchange.

The following sources of data were used to calculate the measure:

1. QHP products: Claims data from issuers, consisting of hospital and office visits, pharmacy, and laboratory claims (when available); enrollment data; and members' demographic data OR

2. Medicare: Claims data from Medicare Parts A, B and D consisting of inpatient and outpatient claims and prescription drug events; enrollment data; and beneficiaries' demographic data.

Please note that Medicare data were used to supplement QHP data for measure testing because they offer a robust sample for calculation of measure performance reliability. Medicare PDPs are similar to QHPs in that they are offered by private insurance companies and are responsible for providing safe and effective medication management. Additionally, if variation in performance is similar among QHP products and Medicare PDPs, we could conclude this measure is generally applicable and reliable at the health plan level. At the time this form was completed, CMS does not have a plan to add this measure to quality reporting or value-based purchasing programs for Medicare enrollees but may consider this measure for the future.

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)*

Health Plan

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

Outpatient Services

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 because this is not a composite performance measure.

2. Validity – See attached Measure Testing Submission Form

AMO_NQF_Testing_Attachment_20190801_FV-637076083655471242.docx

2.1 For maintenance of endorsement

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

2.2 For maintenance of endorsement

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

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Annual Monitoring for Persons on Long-Term Opioid Therapy
Date of Submission: 10/28/2019

Type of Measure:

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

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*).

The measured entities include Qualified Health Plans (QHPs) that operate in the Health Insurance Exchanges, established by the Patient Protection and Affordable Care Act. As a condition of participation, QHPs are required to display quality ratings to assist in consumer selection. CMS intends to display global ratings and three summary indicator ratings in future years. The measure described in this form is being considered for inclusion for this program, known as the Quality Rating System (QRS). To test the measure in QHPs, the following data were used:

- 2015–2016 administrative claims data from four issuers (referred to as QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4), containing a total of seven Health Insurance Exchange Qualified Health Plan (QHP) products in 2015 and eight in 2016. The following describes the terminology of the units associated with the Health Insurance Exchange: "Issuer" refers to an individual insurance company or insurance organization. The term "product" refers to a package of health coverage benefits that are offered using a particular network type (i.e., health maintenance organization, preferred provider organization, exclusive provider organization, point of service, or indemnity).[1] Unique products for each issuer are referred to using alphabetic labeling (e.g., two unique products from the same issuer are referred to as Product A and Product B).

Please note that Medicare data were used for measure testing to enhance the measure testing results because they offer a robust sample for measure testing, such as calculation of measure performance reliability. Medicare PDPs are similar to QHPs in that they are offered by private insurance companies and are responsible for providing safe and effective medication management. Additionally, if variation in performance is similar among QHP products and Medicare PDPs, we could conclude this measure is generally applicable and reliable at the health plan level. At the time this form was completed, CMS does not have a plan to add this measure to quality reporting or value-based purchasing programs for Medicare enrollees but may consider this measure for the future. However, this measure is being considered for use in the Quality Rating System for Qualified Health Plans. Qualified Health Plans (QHPs) operate in the Health Insurance Exchanges, established by the Patient Protection and Affordable Care Act. As a condition of participation, eligible QHPs are required to collect and submit quality measure data. CMS calculates quality ratings based on the data submitted and Exchanges are required to display QHP overall quality ratings and three summary indicator ratings to assist in consumer selection of a QHP offered on an Exchange.

To test the measure in PDPs, the following data were used:

- 2015–2016 administrative claims data from Medicare Parts A, B, and D for beneficiaries enrolled in stand-alone Part D Prescription Drug Plans (referred to as Medicare PDPs)

Citation:

1. Centers for Medicare & Medicaid Services. Federal Definitions for Health Insurance Products and Plans. Baltimore, MD: US Department of Health and Human Services; 2016. https://www.cms.gov/CCIIO/Resources/Training-Resources/Downloads/product-vs-plan-ppt.pdf. Accessed June 12, 2018.

1.3. What are the dates of the data used in testing? January 1, 2015 – December 31, 2016

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*)

The measured entities include Qualified Health Plans (QHPs) operate in the Health Insurance Exchanges, established by the Patient Protection and Affordable Care Act. As a condition of participation, QHPs are required to display quality ratings to assist in consumer selection. CMS intends to display global ratings and three summary indicator ratings in future years. The measure described in this form is being considered for inclusion for this program, known as the Quality Rating System (QRS).

Characteristics of the data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are summarized in Tables 1.5 a (2015) and 1.5 b (2016). The data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4 included all members with claims associated with the QHP products. In order for the testing to be reflective of how the measure would perform if it were implemented in the QRS, testing methodology aligned with the 2018 Quality Rating System Measure Technical Specifications which can be found at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf[1]

- QHP products with 500 or fewer total members were excluded from all analyses, and
- Denominators had to have at least 30 members in order to show the results of analyses.

The 501 member and 30 minimum denominator rules are not part of the measure specifications. The analyses followed these rules to reflect steps that would be taken if the measure were implemented into the Quality Rating System (QHP data).

Medicare data were used for measure testing to enhance the measure testing results. At the time this form was completed, CMS does not have a plan to add this measure to quality reporting or value-based purchasing programs for Medicare enrollees but may consider this measure for the future. The Medicare sample included all beneficiaries from the national Medicare claims database who had at least one month of Part A and Part B coverage and no HMO coverage during the year and who were in a stand-alone Medicare PDP. The 501 member and 30 minimum denominator rules were not applied to the Medicare data since the rules are specific to the Quality Rating System (QHP data).

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-

Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

Table 1.5 a. 2015 Sample Characteristics of the Data

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Number of QHP Products or Medicare PDPs	3	1	2	1	67
Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP	289,136	49,137	15,671	3,354	18,257,146
Mean # of Members/ Beneficiaries per Product/PDP	96,378	49,137	7,836	3,354	272,495

Table 1.5 b. 2016 Sample Characteristics of the Data

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Number of QHP Products or Medicare PDPs	3	1	3	1	63
Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP	223,427	33,205	84,255	2,284	18,945,015
Mean # of Members/ Beneficiaries per Product/PDP	74,476	33,205	28,085	2,284	300,715

1.6. How many and which <u>patients</u> 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*)

Demographic characteristics of members of QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are shown in Tables 1.6 a (2015) and 1.6 b (2016); however, limited demographic variables were available in our testing data. "N/A" in the tables indicates the data were not available.

Table 1.6 a. 2015 Demographic Characteristics of Members of QHP Issuers and Medicare PDPs

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
Total Sample Size	289,136	49,137	15,671	3,354	18,257,146
Sex n (% of Total Sample)*					
Female	150,116 (51.9)	21,399 (43.5)	7,043 (44.9)	1,538 (45.9)	10,071,540 (55.2)
Male	139,020 (48.1)	27,738 (56.5)	8,628 (55.1)	1,816 (54.1)	8,185,606 (44.8)
Age n (% of Total Sample)*					

Characteristics	QHP Issuer 1	QHP Issuer 2	QHP Issuer 3	QHP Issuer 4	Medicare PDPs
<18 years	9,584 (3.3)	3,600 (7.3)	1,578 (10.1)	247 (7.4)	111 (0.0)
18–26 years	38,590 (13.4)	3,633 (7.4)	1,640 (10.5)	333 (9.9)	89,804 (0.5)
27–44 years	81,098 (28.0)	12,486 (25.4)	5,671 (36.2)	1,022 (30.5)	864,242 (4.7)
45–64 years	152,252 (52.7)	28,965 (59.0)	6,603 (42.1)	1,711 (51.0)	2,813,147 (15.4)
≥65 years	7,612 (2.6)	453 (0.9)	179 (1.1)	41 (1.2)	14,489,842 (79.4)
Race n (% of Total Sample)*					
White/ Caucasian	N/A	N/A	N/A	N/A	15,275,375 (83.7)
African-American	N/A	N/A	N/A	N/A	1,826,519 (10.0)
Hispanic	N/A	N/A	N/A	N/A	368,352 (2.0)
Other	N/A	N/A	N/A	N/A	608,822 (3.3)
Unknown	N/A	N/A	N/A	N/A	178,078 (1.0)

*Numbers in parentheses represent the column percent by demographic characteristic.

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

The following sources of data were used in testing the measure at the health plan level:

1. QHP products - claims data from issuers, consisting of hospital and office visit, pharmacy, and laboratory claims (when available); enrollment data; members' demographic data; and provider information.
2. Medicare - claims data from Medicare Parts A and B and stand-alone Part D PDPs, consisting of inpatient and outpatient claims and prescription drug events; enrollment data; members' demographic data; and provider information.

Table 1.7. Data Used to Test the Measure

Aspects of Testing	QHP Data	Medicare Data
Development of the Denominator	✓	✓
Development of the Numerator	✓	✓
Data Element Feasibility	✓	✓

Aspects of Testing	QHP Data	Medicare Data
Measure Performance Reliability (Signal to Noise)	✓	✓
Calculating Measure Performance	✓	✓
Exclusion Analyses	✓	✓
Disparities Analyses	✓	✓

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.

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*)

Measure score reliability was estimated using a beta-binomial model based on the signal-to-noise approach, which determines how well performance can be distinguished between products. The signal-to-noise ratio was calculated as a function of the variance between products (signal) and the variance within a product (noise).[1] Reliability estimates for Medicare PDPs were computed by using the methods of minimum denominator and volume categories, described by Scholle et al. (2008).[2] For the QHP data, the mean reliability score was calculated across QHP products. This difference in approach is due to the limited number of available QHP products.

Reliability scores can range from 0.0 to 1.0. A score of 0.0 implies that all variation is completely attributable to measurement error (i.e., noise or the product variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance across products. Based on Adams' work, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to distinguish differences in performance.[1]

Citations:

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

2. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care.* 2008;14(12):833-838.

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, QHP Products, Issuer 1, Issuer 2, Issuer 3, and Issuer 4

Among the seven QHP products tested, reliability ranged from 0.59 to 0.99 with a mean reliability score of 0.85 indicating the measure can reliably distinguish performance between QHP products (Table 2a2.3 a).

Table 2a2.3 a. 2016 Reliability Among QHP Products with At Least 30 Members in the Denominator

QHP Issuer	Product	Denominator	Numerator	Measure Rate	Variance Within	Variance Between	Reliability Score
1	A	49	40	81.6%	30.60	97.93	0.76
1	B	1,299	1,040	80.1%	1.23	97.93	0.99
1	C	47	39	83.0%	30.05	97.93	0.76
2	A	702	437	62.3%	3.35	97.93	0.97
3	A	462	268	58.0%	5.27	97.93	0.95
3	B	391	233	59.6%	6.16	97.93	0.94
4	A	36	21	58.3%	67.51	97.93	0.59
Mean							**0.85**

Minimum Denominator for Reliability, Medicare PDPs

Using the method of minimum denominator and volume categories, a minimum of 100 beneficiaries in the denominator results in an overall reliability score of 0.72, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between units of analysis. Of the 62 PDPs with at least one patient in the denominator, the majority (83.9%) of PDPs had at least 100 beneficiaries in the measure denominator representing a mean performance rate of 67.3% (reliability = 0.72) (Table 2a2.3 b).

Table 2a2.3 b. 2016 Medicare PDP Reliability and Assessment of Adequacy for Tests Conducted

Min Denominator	Total # of PDPs	# of PDPs with at Least 100 Beneficiaries	Mean Rate of Plans with at Least 100 Beneficiaries	Reliability Score
100	62	52	67.3%	0.72

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?*)

Reliability, QHP Products, Issuer 1, Issuer 2, Issuer 3, and Issuer 4

The results indicate that the measure is reliable at the health plan level, based on a sample of QHP products. Among the products with at least 30 denominator members, the average reliability score was 0.85, which suggests sufficient signal relative to noise to discriminate performance between plans.

Reliability, Medicare PDPs

The results indicate that the measure is reliable at the health plan level, based on Medicare PDP data with at least 100 beneficiaries in the denominator. In 2016, the majority of Medicare PDPs (83.9%) had at least 100 beneficiaries in the denominator, which produced measure performance rates with sufficient reliability (0.72) to distinguish differences in performance among plans. A larger denominator threshold was needed for

reliable scores using the PDP data for two reasons. First, there was more observed variation among products in the QHP population than in the Medicare PDP population. This increases the ability to discriminate among products and lowers the minimum denominator. Second, given the small number of QHP products, we could not use the standard method of minimum denominator and volume categories, since this method involves estimating variation among products in subsets of the original dataset. Instead we looked all products large enough to report in the QHP and found the mean reliability was above the threshold of 0.7. This suggests that, on average, products with at least 30 denominator members are reliable. Using the more stringent method of minimum denominator and volume categories might lead to a larger minimum denominator for QHPs, if there were enough data to do the analysis.

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)*

Face validity:

HSAG's Technical Expert Panel (TEP) evaluated the face validity of the measure and the measure score after testing was completed. The TEP is composed of three representatives from large QHP issuers and nine representatives from other stakeholder groups, such as measurement industry representatives, clinical and nonclinical experts, and patient/caregiver representatives. The evaluation of face validity was conducted through a webinar meeting with voting functions. TEP members were specifically asked whether they agree with the following statement: "The performance scores resulting from the measure *Annual Monitoring for Persons on Long-Term Opioid Therapy (AMO)*, as specified, can be used to distinguish good from poor plan-level quality related to the process of administrating at least one drug test during the measurement year among those with long-term opioid therapy." They responded on a "yes"/"no" scale, indicating either they did or they did not agree with the previous statement.

Validity of ICD-9-CM to ICD-10-CM conversion:

- Goal was to convert the 2015 ICD-9-CM codes used to identify the cancer exclusion to 2016 ICD-10-CM codes and remain consistent with the intent of the measure. Cancer is the only data element in the measure that uses ICD-9-CM or ICD-10-CM codes.
- The full list of codes, with code definitions, is included in the attached excel file.
- Description of the process used to identify ICD-10-CM codes:
    - Names and credentials of experts who assisted in the process:
        - Kimberly Smuk, RHIA (staff)
        - Graves T. Owen, MD (Subject Matter Expert)
        - Paul Jannetto, PhD (Subject Matter Expert)
        - Kara McVey, CPC, CPMA (Subject Matter Expert)

- Name of the tool used to identify/map to ICD-10-CM: CMS GEMs were used in converting the ICD-9-CM code list into ICD-10-CM
- Summary of stakeholder comments received: The Subject Matter Experts reviewed the codes and agreed that the converted ICD-10-CM codes were consistent with the intent of the measure.
- Process:
  1. Started with the target ICD-9-CM code list
  2. Mapped the ICD-9-CM codes using the 2015 CMS GEMs to 2015 ICD-10-CM
  3. Reverse mapped the 2015 ICD-9-CM codes in the CMS GEMs
  4. Manual review of 2015 ICD-9-CM for using key term search
  5. Manual review of relevant ICD-10-CM chapters
  6. Performed evaluation of all codes resulting from above steps – at the staff level and Subject Matter Expert level.

Threats to validity:

This measure requires the specification of drug tests for the numerator calculation. Testing data covers performance years 2015-2016. To identify qualifying drug tests, the measure specifies HCPCS, CPT, and LOINC codes. Prior to 2016, all three code sets allow for the identification of testing for specific drug classes of interest. In 2016, the HCPCS drug test codes were consolidated from 28 drug class-specific codes to four generic codes that differ only by the number of drug classes being tested. This introduced a potential threat to validity, since we were unable to specify drug tests of interest, for particular drug classes, using HCPCS codes in performance years beyond 2015. To test the validity of the measure using these new generic codes, we ran a sensitivity analysis to determine the impact of the 2016 HCPCS changes. The intent of the analysis was to determine if measure performance in 2015 was affected by not being able to identify only drug classes targeted in the measure.

To perform the sensitivity analysis, the Measure Developer calculated two measure rates for 2015, using Issuer 1 and Medicare PDP data. Calculation instructions were given to an analytic firm and rates for Issuer 2, Issuer 3, and Issuer 4 were provided to the Measure Developer.

1. Reference rate: measure rates using the 2015 HCPCS, CPT, and LOINC codes specific to drug classes identified in AACC Academy guidelines for routine screening.[1] The specific rates represent the ability to specify drug classes to test within the 2015 HCPCS coding.
2. Comparison rate: measure rates using 2015 CPT and LOINC codes specific to the drug classes of interest, along with all 2015 HCPCS codes (regardless of type of drug/drug class). The comparison rates represent the effect the HCPCS drug test coding change would have on measure rates in 2016 and beyond.

Citation:

1. Langman L, Jannetto P. *Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients.* The American Association for Clinical Chemistry Academy; 2017. https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients. Accessed May 3, 2018.

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

Systematic Assessment of Face validity:

9 of 9 TEP members (100%) that participated in the voting indicated that the measure has face validity and agreed that the measure was valid as specified.

Validity of ICD-9-CM to ICD-10-CM conversion:

The Subject Matter Experts reviewed the codes and agreed that the converted ICD-10-CM codes were consistent with the intent of the measure.

Threats to validity:

The sensitivity analysis found the effect of the HCPCS code change was negligible. Potential false positives due to inclusion of generic codes that do not specify what drug is tested had little effect on measure performance. As shown in Table 5, the results for Issuer 1 found the Product A measure rate did not change, and the Product B measure rate decreased (improved) 0.1%, a non-statistically significant difference ($p = 0.9180$). Product C did not have ≥30 members in the denominator in 2015, therefore results have been suppressed in alignment with QRS requirements.[1] Results for Issuer 2 show a non-statistically significant, 1.0% decrease (improvement) in the Product A measure rate ($p = 0.3273$). Measure rates did not change in any products for Issuer 3 and Issuer 4. The results among pooled Medicare PDPs produced a 0.5% decrease (improvement) in national measure rates and a 0.2% decrease (improvement) in the mean national rate across PDPs, a non-statistically significant difference ($p = 0.3060$).

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf. Accessed July 13, 2018.

Table 2b1.3 HCPCS Code Change Comparison – Issuer 1 and Medicare 2015 Measure Rates

QHP Issuer and Product	Number of Members Without Drug Test: Numerator/Denominator	Percent Without Drug Test
Issuer 1 Product A (Reference)	33/45	73.3%
Issuer 1 Product A (Comparison)	33/45	73.3%
Issuer 1 Product B (Reference)	1,187/1,411	84.1%
Issuer 1 Product B (Comparison)	1,185/1,411	84.0%
Issuer 1 Product C (Reference)	--	--
Issuer 1 Product C (Comparison)	--	--
Issuer 2 Product A (Reference)	655/944	69.4%
Issuer 2 Product A (Comparison)	646/944	68.4%
Issuer 3 Product A (Reference)	53/74	71.6%
Issuer 3 Product A (Comparison)	53/74	71.6%
Issuer 3 Product B (Reference)	22/38	57.9%
Issuer 3 Product B (Comparison)	22/38	57.9%
Issuer 4 Product A (Reference)	37/61	60.7%
Issuer 4 Product A (Comparison)	37/61	60.7%
Medicare (Reference)	814,584/1,271,123	64.1%
Medicare (Comparison)	808,081/1,271,123	63.6%

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?*)

Face validity:

Of the TEP members who participated in the meeting, 100% of TEP members (9/9) agreed that the measure can be used to distinguish good from poor plan-level quality of care related to the patients on long-term opioid therapy indicating the measure is valid.

Validity of ICD-9-CM to ICD-10-CM conversion:

The Subject Matter Experts reviewed the codes and agreed that the converted ICD-10-CM codes were consistent with the intent of the measure.

Threats to validity:

Findings suggested that when drug tests occurred, they seldom were billed using HCPCS codes and the effect of HCPCS code changes had a non-statistically significant impact on measure rates. Therefore, there is no threat to validity.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

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*)

Individuals with hospice and/or cancer diagnosis (except non-melanoma skin cancer) are excluded. Clinical practice guidelines for prescribing opioids routinely define their scope as outpatient treatment of patients with chronic pain, not including patients with active cancer, palliative care, or end-of-life care. The literature examined aligns with this scope, universally excluding cancer patients. Further, the CMS Opioid Misuse Strategy 2016 does not address treatment of patients with cancer or hospice care, based on the clinical practice guidelines and current literature.[1] The rationale is that such patients require case-by-case decisions made by providers that are based on therapeutic goals and ethical considerations. These exclusions also align with four other opioid quality measures, which exclude hospice and all cancer patients:

- NQF 2940 Use of Opioids at High Dosage in Persons Without Cancer
- NQF 2950 Use of Opioids from Multiple Providers in Persons Without Cancer
- NQF 2951 Use of Opioids from Multiple Providers and at High Dosage in Persons Without Cancer
- NQF 3389 Concurrent Use of Opioids and Benzodiazepines

Individuals with non-melanoma skin cancer are not excluded, which is in alignment with the above-listed measures, opioid misuse risk literature[2,3], and recommendations from the Subject Matter Experts advising the development of this measure.

To determine the effect of the exclusions on the measure rates, measure rates with and without each exclusion were calculated and compared using 2016 data from pooled QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and measure rates from Medicare PDPs.

Citation:

1. Centers for Medicare & Medicaid Services. Opioid Misuse Strategy 2016. Baltimore, MD: US Department of Health and Human Services; 2016. https://www.cms.gov/Outreach-and-Education/Outreach/Partnerships/Downloads/CMS-Opioid-Misuse-Strategy-2016.pdf. Accessed May 8, 2018.

2. Edlund MJ, Martin BC, Fan MY, Devries A, Braden JB, Sullivan MD. Risks for opioid abuse and dependence among recipients of chronic opioid therapy: results from the TROUP study. Drug Alcohol Depend. 2010;112(1-2):90-98. doi: 10.1016/j.drugalcdep.2010.05.017.

3. Sullivan MD, Edlund MJ, Fan MY, Devries A, Brennan Braden J, Martin BC. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and medicaid insurance plans: The TROUP Study. Pain. 2010;150(2):332-339. doi: 10.1016/j.pain.2010.05.020.

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*)

To determine the effect of the exclusions on the 2016 measure rates, the pooled rates were calculated with and without each exclusion, as shown in Table 2b2.2.

Table 2b2.2 2016 AMO Measure Rate by Exclusion Status

	Denominator	Numerator	Measure Rate	95% CI
QHP Issuer 1				
No exclusions	1,670	1,341	80.3%	(78.4, 82.2)
Cancer excluded	1,395	1,119	80.2%	(78.1, 82.3)
Hospice excluded	1,664	1,336	80.3%	(78.4, 82.2)
Cancer and Hospice excluded	1,395	1,119	80.2%	(78.1, 82.3)
QHP Issuer 2				
No exclusions	924	592	64.1%	(61.0, 67.2)
Cancer excluded	702	437	62.3%	(58.7, 65.8)
Hospice excluded	920	588	63.9%	(60.8, 67.0)
Cancer and Hospice excluded	702	437	62.3%	(58.7, 65.8)
QHP Issuer 3				
No exclusions	1,173	700	59.7%	(56.8, 62.5)
Cancer excluded	881	517	58.7%	(55.4, 62.0)
Hospice excluded	1,162	690	59.4%	(56.5, 62.3)
Cancer and Hospice excluded	880	517	58.8%	(55.4, 62.1)
QHP Issuer 4				
No exclusions	44	24	54.5%	(39.8, 69.3)
Cancer excluded	36	21	58.3%	(42.2, 74.4)
Hospice excluded	44	24	54.5%	(39.8, 69.3)
Cancer and Hospice excluded	36	21	58.3%	(42.2, 74.4)
Medicare PDPs				
No exclusions	1,818,757	1,130,967	62.2%	(62.1, 62.3)

Cancer excluded	1,260,496	764,234	60.6%	(60.6, 60.7)
Hospice excluded	1,766,315	1,086,447	61.5%	(61.4, 61.6)
Cancer and Hospice excluded	1,235,466	742,632	60.1%	(60.0, 60.2)

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*)

As shown in Table 2b2.2, the impact on measure rates is minimal from excluding patients with cancer and/or those receiving hospice care. However, these exclusions will be retained in the measure specifications to align with other NQF-endorsed measures, clinical practices guidelines, and expert recommendations.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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 not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable.

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

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.**

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

Not applicable.

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

Not applicable.

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

Not applicable.

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

Not applicable.

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

Not applicable.

2b3.9. Results of Risk Stratification Analysis:

Not applicable.

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

Not applicable.

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

Not applicable.

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

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

To identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the prescription drug plan and product level with data from 2016. Confidence intervals (CI 95%) were calculated around point estimates for each plan and product and then compared to the respective overall mean of prescription drug plans and product levels. If the confidence intervals did not overlap with the overall mean, the difference was 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*)

Meaningful Differences at the Plan Level – 2016

Measure rates across QHP products ranged from 58.0% to 83.0% (Table 2b4.2 a) with a mean measure rate of 69.0%. These rates were substantially higher (indicating worse performance) than the rate observed among Medicare PDPs (67.3%) (Table 2b4.2 b). The confidence intervals for the QHP product measure rates did not overlap with the overall mean, indicating that the differences were considered statistically significant.

Table 2b4.2 a 2016 QHP Performance for Those with at Least 30 Members in the Denominator

QHP Issuer	QHP Product	Rate	Confidence Interval
1	A	81.6%	69.8, 93.5
1	B	80.1%	77.9, 82.3
1	C	83.0%	71.2, 94.8
2	A	62.3%	58.7, 65.8
3	A	58.0%	53.4, 62.6
3	B	59.6%	54.6, 64.6
3	C	-	-
4	A	58.3%	42.2, 74.4

Based on the reliability findings for Medicare PDPs, we calculated variation among PDPs with at least 100 denominator beneficiaries. We found that 36.5% (19/50) of PDPs had rates significantly lower than the mean, and 42.3% (22/52) of PDPs had rates significantly greater than the mean, based on the confidence interval calculations. The difference in performance between high-performing (i.e., 10th percentile) and low-

performing (i.e., 90th percentile) PDPs was 20.1%, indicating both variation between high- and low-performing PDPs and suboptimal performance across PDPs (Table 2b4.2 b).

Table 2b4.2 b. 2016 Medicare PDP Performance for Those with at Least 100 Beneficiaries in the Denominator

n Plans	Mean	Minimum	P10	P25	P50	P75	P90	Maximum
52	67.3%	48.3%	55.8%	62.8%	68.7%	73.1%	75.9%	80.7%

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 low performance rates of the QHP products (average rate of 69.0% in 2016) suggests substantial opportunity for improvement in the management of individuals on long-term opioid therapy among QHPs in the Health Insurance Exchanges. In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 67.3%) among Medicare PDPs. The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries on long-term opioid therapy.

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*)

This is a claims-based measure and relies on final paid claims from payors (Medicare, QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, or QHP Issuer 4). Missing data was not an issue for measure calculation.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Not applicable.

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

Not applicable.

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

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 claims

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

This measure is specified using administrative claims data. At this time, there is no plan to specify the measure as an 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:

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.

Testing demonstrated that the measure was feasible to be specified and calculated using administrative claims data from QHP products and Medicare PDPs. Data used in the calculation of this measure are obtained from administrative claims, which are routinely, reliably, and securely collected for billing purposes. We do not anticipate any feasibility or implementation issues related to data collection for this measure. No threats to the validity of this measure were identified using a limited analysis designed to address missing data.

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)*.

PQA retains the rights of ownership to this measure, and can rescind or alter the measure at any time. No use of any PQA measure is authorized without prior PQA approval of such use. All uses of PQA measures are subject to such conditions as PQA specifies, and certain uses of the measures may be subject to a licensing agreement specifying the terms of use and the licensing fee. Users of the measure shall not have the right to alter, enhance, or otherwise modify the measures.

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 Quality Improvement (Internal to the specific organization)	

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 being submitted for initial endorsement.

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?*)

Not applicable; the measure is being submitted for initial endorsement.

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.*)

In the Final 2019 Call Letter for the Quality Rating System (QRS) and Qualified Health Plan Enrollee Experience Survey (QHP Enrollee Survey), CMS announced that it will pursue the measure Annual Monitoring for Patients [Persons] on Chronic [Long-Term] Opioid Therapy for inclusion in the QRS measure set because of the potential for this measure to reduce harm in the delivery of care, and because of the important policy priorities it addresses.[1] CMS anticipates proposing to add the measure beginning with the 2021 QRS as part of the 2020 QRS Call Letter process with data collection beginning with the 2021 ratings year, and CMS scoring in 2022.

Citation:

1. Centers for Medicare & Medicaid Services. Health Insurance Exchange. Final 2019 Call Letter for the Quality Rating System (QRS) and Qualified Health Plan Enrollee Experience Survey (QHP Enrollee Survey). Finalized QRS and QHP Enrollee Survey Program Refinements. June 2019. Baltimore, MD: US Department of Health and Human Services; 2019. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2019_Call_Letter_for_QRS_and_QHP_Enrollee_Experience_Survey_508.pdf. Accessed on: 10/18/2019.

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.

The measure is not currently implemented in a public reporting program, and therefore, there is no information available regarding feedback during implementation. The Technical Expert Panel (TEP) reviewed the updated measure evidence, testing, and performance results and interpretation via several webinar conferences. The TEP was comprised of three representatives from large issuers, and nine from other stakeholder groups, including academic institutions, quality organizations, government agencies, and patient/caregiver representatives. In addition to the TEP, three Subject Matter Experts (SMEs) reviewed the measure specifications, and performance results and interpretation, via several webinar conferences. Additionally, feedback was solicited, received, and addressed from the Centers for Disease Control and

Prevention. A full list of TEP members and SMEs, and their organizations, is noted in the Additional Information section.

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.

Meetings with the TEP were held throughout 2015-2017. During these meetings, TEP members were provided with updated measure evidence, development, and testing results. Data necessary to judge the validity and usability of the measure were provided, along with the measure algorithm and a complete list of codes used to calculate the measure. Questions that arose from these meetings were addressed either during the meeting or in follow-up communications.

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.

TEP members were encouraged to provide feedback throughout the measure development process by means of meeting discussions, voting, and through follow-up communications. Members were asked whether they thought the measure has clear and precise specifications and face validity, as defined by NQF, to distinguish good from poor plan-level quality of care related to individuals on long-term opioid therapy.

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

There were three TEP members that represented health plans. They unanimously agreed that the measure has clear and precise specifications and can distinguish good from poor plan-level quality related to the process of administering at least one drug test during the year among those with long-term opioid therapy (i.e., the measure has face validity).

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

There were nine TEP members who represented other stakeholder groups. They voted to indicate whether this measure has clear and precise specifications and can distinguish good from poor plan-level quality of care related to individuals on long-term opioid therapy. Of the nine members, three members did not participate in the vote. The six members that did respond unanimously agreed that the measure has clear and precise specifications and can distinguish good from poor plan-level quality related to the process of administering at least one drug test during the year among those with active long-term opioid therapy (i.e., the measure has face validity). All feedback received from the TEP regarding this measure indicates that the measure will be useful for health plans to improve quality of care individuals on long-term opioid therapy and can be implemented without undue burden for reporting for the Quality Rating System.

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.

TEP and SME feedback was considered throughout the measure development process. Feedback that was considered during the measure development process includes:

1. Support to include all individuals on long-term opioid therapy in the measure, rather than limiting the measure to new long-term therapy.

2. Recommendations on the list of opioid formulations, indications, and routes to include in the denominator population. These recommendations led to the inclusion of all opioid formulations indicated for pain, and the exclusion of opioids indicated for intravenous or epidural routes.

3. Recommendations to allow for any drug test throughout the measurement year to count towards the numerator calculation.

4. Recommendations on what drug tests to specify in the numerator. These recommendations included support for aligning with the routine monitoring list in the American Association for Clinical Chemistry (AACC) Academy guidelines,[1] as well as specific drug classes from the AACC Academy's high-risk monitoring list. The

team decided to align with the routine monitoring list only, to encourage a minimum standard of care, which includes testing for opioids.

5. Recommendation to use the term "long-term opioid therapy" rather than "chronic opioid therapy" given the use of the terms "chronic pain" and "acute pain," and that using the term chronic in conjunction with length of opioid therapy can be confusing and misapplied. This recommendation resulted in the title of the measure being modified.

Citations

1. Langman L, Jannetto P. Laboratory Medicine Practice Guidelines: Using Clinical Laboratory Tests to Monitor Drug Therapy in Pain Management Patients. The American Association for Clinical Chemistry Academy; 2017. https://www.aacc.org/science-and-practice/practice-guidelines/using-clinical-laboratory-tests-to-monitor-drug-therapy-in-pain-management-patients. Accessed May 3, 2018.

Improvement

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

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

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

This measure is actionable by both providers and plans. Evidence suggests correlations between drug tests and subsequent changes in the rates of prescriptions, risk for ADEs, and clinical follow-up.[1-3] One study found that clinicians planned to change the treatment plan in 69% of 83 cases in which a patient tested positive for aberrant behavior determined through the use of drug testing.[1] Rates of drug testing are suboptimal among individuals on long-term opioid therapy and were reported as low as 8%.[4] Additionally, variation in rates of drug testing for individuals on long-term opioid therapy exists between similar sites of care (i.e., three safety-net primary care clinics: 31%, 35%, and 67%).[5]

However, the literature suggests that multicomponent education and resource interventions for physicians can improve guideline concordant care. In one study, at one-year post-intervention, patients with an intervention physician were more likely than those with control physicians to receive guideline-concordant care (65.9% versus 37.8%; p <.001).[6] Specific to urine drug testing, patients were more likely to have received at least one drug test within one year if they had an intervention physician, compared to a control physician (74.6% versus 57.9%; p <.001).

The results of drug tests are critical sources of information for providers serving patients receiving long-term opioid therapy. The measure will assist clinicians by encouraging evidence-based processes that enhance patient safety for patients on long-term opioid therapy. These processes will help clinicians to identify patients on long-term opioid therapy who engage in aberrant drug-related behaviors and can help to identify patients who need referral for opioid disorder. Ultimately, the measure should help to lower the risk of ADEs, including substance abuse and drug-related mortality, for patients who are on long-term opioid therapy. This is supported by a recent study which found that, independent of patient-level risk factors, higher average levels of administered drug tests within VA facilities predicted a significant reduction in risk of prescription opioid-related suicide and overdose events. This study further suggests that for every additional one percent of opioid-prescribed patients that receive drug test monitoring, there is up to a one percent reduction in patient risk of an opioid-related suicide or overdose event.[7]

Citations

1. Morasco BJ, Krebs EE, Cavanagh R, Hyde S, Crain A, Dobscha SK. Treatment changes following aberrant urine drug test results for patients prescribed chronic opioid therapy. Journal of opioid management. 2015;11(1):45-51. doi: 10.5055/jom.2015.0251.

2. Worley J. Prescription drug monitoring programs, a response to doctor shopping: purpose, effectiveness, and directions for future research. Issues in mental health nursing. 2012;33(5):319-328. doi: 10.3109/01612840.2011.654046.

3. Baehren DF, Marco CA, Droz DE, Sinha S, Callan EM, Akpunonu P. A statewide prescription monitoring program affects emergency department prescribing behaviors. Annals of emergency medicine. 2010;56(1):19-23 e11-13. doi: 10.1016/j.annemergmed.2009.12.011.

4. Starrels JL, Becker WC, Weiner MG, Li X, Heo M, Turner BJ. Low use of opioid risk reduction strategies in primary care even for high risk patients with chronic pain. J Gen Intern Med. 2011;26(9):958-964. doi: 10.1007/s11606-011-1648-2.

5. Lange A, Lasser KE, Xuan Z, et al. Variability in opioid prescription monitoring and evidence of aberrant medication taking behaviors in urban safety-net clinics. Pain. 2015;156(2):335-340. doi: 10.1097/01.j.pain.0000460314.73358.ff.

6. Liebschutz JM, Xuan Z, Shanahan CW, et al. Improving Adherence to Long-term Opioid Therapy Guidelines to Reduce Opioid Misuse in Primary Care: A Cluster-Randomized Clinical Trial. JAMA internal medicine. 2017;177(9):1265-1272. doi: 10.1001/jamainternmed.2017.2468.

7. Brennan PL, Del Re AC, Henderson PT, Trafton JA. Healthcare system-wide implementation of opioid-safety guideline recommendations: the case of urine drug screening and opioid-patient suicide- and overdose-related events in the Veterans Health Administration. Transl Behav Med. 2016;6(4):605-612. doi: 10.1007/s13142-016-0423-7.

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.

This is a new measure and has not been in use, and therefore there are no unexpected findings from implementation to report.

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

This is a new measure and has not been in use, and therefore there are no unexpected benefits from implementation to report.

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)

1617 : Patients Treated with an Opioid who are Given a Bowel Regimen

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

3316 : Safe Use of Opioids – Concurrent Prescribing

3389 : Concurrent Use of Opioids and Benzodiazepines (COB)

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

Risk of Continued Opioid Use (HEDIS measure stewarded by NCQA)

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.

An environmental scan revealed related measures listed above, which share similar populations of interest (patients receiving opioids). NQF 1617 targets vulnerable adults given a new prescription for an opioid, and therefore has a different target population than the AMO measure. NQF 3316e is an eCQM that targets patients discharged from a hospital-based encounter, a different setting of care than the AMO measure. Harmonization of value sets has been addressed to the extent possible with related outpatient health plan measures, NQF 2940, 2950, 2951, and 3389, including the cancer and hospice exclusions and targeted opioid medications. The AMO measure's area of focus (numerator) does not overlap with any existing measure, and its focus on drug tests for patients on long-term opioid therapy is unique. Therefore, while there are some related measures that evaluate similar target populations of patients receiving opioid therapy, the AMO measure is a new and evidence-based focus to empower health plans to address opioid misuse and opioid use disorder, and improve patient safety. Harmonization has been addressed to the extent possible, and PQA will continue to identify and address opportunities to harmonize with related measures over time.

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.

1. Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested

information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

2. Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Pharmacy Quality Alliance

Co.2 Point of Contact: Lynn, Pezzullo, lpezzullo@pqaalliance.org, 401-474-9706-

Co.3 Measure Developer if different from Measure Steward: Health Services Advisory Group

Co.4 Point of Contact: Melissa, Castora-Binkley, mcastora-binkley@hsag.com, 813-865-3182-

3. 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.

Technical Expert Panel (TEP), 2015-2017

1. Andy Amster, MSPH, Kaiser Permanente National Office
2. Marybeth Farquhar, PhD, MSN, RN, URAC
3. Susan Fitzpatrick, RN, BSN, Cigna Healthcare
4. Aparna Higgins, Duke-Margolis Center for Health Policy; Brandeis University
5. Jon Mark Hirshon, MD, PhD, MPH, University of Maryland, School of Medicine
6. Christine Hunter, MD, US Office of Personnel Management
7. Carol Keegan, PhD, Patient representative
8. Dana Mukamel, PhD, University of California, Irvine
9. Chinwe Nwosu, NS, America's Health Insurance Plans
10. Derek Robinson, MD, MBA, FACEP, Health Care Service Corporation
11. Arlene Salamendra, Patient representative
12. Ted von Glahn, MSPH, von Glahn Consulting

Workgroup, 2015-2017

1. Andy Amster, MSPH, Kaiser Permanente National Office
2. Marybeth Farquhar, PhD, MSN, RN, URAC
3. Susan Fitzpatrick, RN, BSN, Cigna Healthcare
4. Jon Mark Hirshon, MD, PhD, MPH, University of Maryland, School of Medicine
5. Dana Mukamel, PhD, University of California, Irvine
6. Derek Robinson, MD, MBA, FACEP, Health Care Service Corporation
7. Arlene Salamendra, Patient representative
8. Ted von Glahn, MSPH, von Glahn Consulting

The TEP evaluated this process measure, drafted by Health Services Advisory Group (HSAG), in regard to the four primary measure evaluation criteria used in the NQF consensus endorsement process (importance,

scientific acceptability, feasibility, and use/usability). The TEP discussed the strengths and weaknesses of the proposed measure and made recommendations regarding measure specifications, and inclusion and exclusion criteria.

The workgroup, a subgroup of the TEP, reviewed the measure specifications in regard to the four primary measure evaluation criteria used in the NQF consensus endorsement process (importance, scientific acceptability, feasibility, and use/usability). The workgroup discussed the strengths and weaknesses of the proposed measure and made recommendations regarding measure specifications, and inclusion and exclusion criteria.

Subject Matter Experts (SMEs), 2015-2017

1. Paul Jannetto, PhD, Mayo Clinic
2. Graves T Owen, MD, Texas Pain Rehabilitation Institute
3. Kara McVey, CPC, CPMA, ILEX Consulting LLC

SMEs reviewed the measure specifications regarding scientific and administrative accuracy, discussed the strengths and weaknesses of the measure and made recommendations regarding measure specifications, and inclusion and exclusion criteria. The SMEs also participated in workgroup and TEP meetings and provided additional information where necessary.

Subject Matter Expert (SME), 2019

Feedback on the measure was sought from the Centers for Disease Control and Prevention. Jan L. Losby, PhD, MSW, Team Lead, Opioid Overdose Health Systems Team, Division of Unintentional Injury Prevention, provided written feedback that was incorporated into the measure specifications.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2019

Ad.3 Month and Year of most recent revision: 10, 2019

Ad.4 What is your frequency for review/update of this measure? Annually.

Ad.5 When is the next scheduled review/update for this measure? 06, 2020

Ad.6 Copyright statement: Rights retained by PQA, Inc 2019.

Limited proprietary coding is contained in the Measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets.

CPT[R] contained in the measure specifications is copyright 2004-2017 American Medical Association.

ICD-10 copyright 2017 World Health Organization. All Rights Reserved.

LOINC[R] copyright 2004-2017 Regenstrief Institute, Inc.

Uniform Bill Codes copyright 2017 American Hospital Association. All rights reserved.

Due to technical limitations, registered trademarks are indicated by (R) or [R].

Ad.7 Disclaimers: This performance measure does not establish a standard of medical care and has not been tested for all potential applications.

Ad.8 Additional Information/Comments: