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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 3475e

Measure Title: Appropriate Use of DXA Scans in Women Under 65 Years Who Do Not Meet the Risk Factor Profile for Osteoporotic Fracture

Measure Steward: Centers for Medicare & Medicaid Services, Center for Clinical Standards and Quality, Quality Measurement and Value-Based Incentives Group (QMVIG), Division of Electronic and Clinician Quality, MS S3-02-01

Brief Description of Measure: Percentage of female patients 50 to 64 years of age without select risk factors for osteoporotic fracture who received an order for a dual-energy x-ray absorptiometry (DXA) scan during the measurement period.

Developer Rationale: This measure is expected to increase the recording of patient risk for fracture data and to decrease the number of inappropriate DXA scans. Current osteoporosis guidelines recommend using bone measurement testing to assess osteoporosis risk in women ages 65 and older. In postmenopausal women younger than 65, guidelines recommend using a formal clinical risk assessment tool to establish patients' risk for osteoporosis in order to determine whether to screen them for osteoporosis using bone measurement testing. Clinical information such as age, BMI, parental history of hip fracture, smoking, and alcohol use can be used to determine a woman's fracture risk (U.S. Preventive Services Task Force, 2018).

In addition, there are potentially avoidable harms associated with screening for osteoporosis in general, including exposure to radiation, false-positive exams, and the side effects of unnecessary osteoporosis medications, which add costs to an already burdened health care system (Lim et al., 2009).

Citations:

Lim LS, Hoeksema LJ, Sherin K. Screening for osteoporosis in the adult U.S. population: ACPM position statement on preventive practice. Am J Prev Med. 2009;36(4):366-75.

U.S. Preventive Services Task Force. Screening for osteoporosis to prevent fractures: U.S. Preventive Services Task Force recommendation statement." JAMA. 2018;319(24):2521-31.

Numerator Statement: Female patients who received an order for at least one DXA scan in the measurement period.

Denominator Statement: Female patients ages 50 to 64 years with an encounter during the measurement period.

Denominator Exclusions: The measure excludes patients who have a combination of risk factors (as determined by age) or one of the independent risk factors.

Measure Type: Process: Appropriate Use

Data Source: Electronic Health Records

Level of Analysis: Clinician: Individual

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 an overuse measure aiming to decrease inappropriate DXA screenings for osteoporosis and reduce avoidable harms associated with screening patients who have a low risk of osteoporotic fractures. The measure is based on the 2018 USPTF guideline, which is based on Grade B evidence: "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in postmenopausal women younger than 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool."
- About 40% of women who have received a DXA scan do not meet risk factors for frailty, and may receive inappropriate medication and treatment for osteoporosis or osteopenia; the developers cite a study that showed that up to two-thirds of newly prescribed osteoporosis medications were given based on abnormalities identified using DXA scans that do not meet clinical guidelines for diagnosis.
- Potential harms caused by overuse of screening for osteoporosis include "false-positive test results, which can lead to unnecessary treatment, and false-negative test results" as well as "radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)."
- An evidence review conducted in 2018 captured 168 published articles of good or fair quality on screening for and treatment of osteoporotic fractures, risk assessment tools, and the efficacy of screening.

Exception to evidence

N/A

Questions for the Committee:

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

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: moderate; Consistency: high (Box 5) → Moderate (Box 5b) → Moderate

The highest possible rating is moderate.

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

RATIONALE:

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.

- In data on 7.5 million women from one large health plan, 6.7 percent of the women ages 50 to 64 had potentially inappropriate DXA scans.
- About 40% of women who have had a DXA scan do not meet the risk factors for frailty.
- A retrospective cohort study of 13 practices assessed the three-, five-, and seven-year incidence of inappropriate and appropriate DXA scans. This study revealed a three-year incidence of DXA scans of 18.4 percent in women ages 50 to 59 without osteoporosis risk factors, and 24.9 percent in women ages 60 to 64 without risk factors.

Disparities

Overuse rates vary by race, with white women and Asian women having higher rates of overuse.

Rates of potentially inappropriate DXA scans by age and race from three test sites – Percents of Scans
(calculated using earlier version of measure for ages 18-64)

Site	White	Black	Asian	Other	Missing
Site 1	0.11	0.07	0.12	0.05	0.08
Site 2	2.36	1.23	2.43	4.87	1.83
Site 3	2.79	2.67	1.76	1.72	2.28

There are also disparities in general use of DXA scans and osteoporosis care:

- A gender matched study on women ages 60 and older in primary care practices, only 29.8 percent of black women were referred for a DXA scan, compared with 38.4 percent of white women. Of the referred women, 20.8 percent of the black women had the scan, compared with 27.0 percent of the white women.
- Among included women with a diagnosis of osteoporosis, black women were less likely to receive medication (79.6 percent) than were white women (89.2 percent) ($p < 0.05$), controlling for both age and BMI. But there was no difference in the pattern of follow-up visits between the two races.
- The prevalence of osteoporosis differs across races and ethnicities. In 2010, an estimated 15.8 percent of non-Hispanic white women, 7.7 of non-Hispanic black women, and 20.4 percent of Mexican American women had osteoporosis of femoral neck or lumbar spine.

Questions for the Committee:

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

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

RATIONALE:

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

1a. Evidence

Comments***:***

** Yes the evidence relates well, is applied directly and is current.

**variation in care, and overuse of testing is shown.

**overuse of imaging in women less than65 without risk factors, and over treatment of women less than 65 with abnormalities not meeting clinical guidelines for treatment. Recent USPTF guideline cited.

**Decreasing unnecessary tests is a desired outcome.

**Inappropriate screening relates indirectly to the adverse outcome of treatment when harm exceeds benefit. I am aware of no new evidence. Inappropriate screening relates indirectly to the adverse outcome of treatment when harm exceeds benefit. I am aware of no new evidence.

**There is moderate evidence to support this overuse measure

**Evidence from literature review and the USPSTF report apply directly to the process of appropriate or inappropriate selection of women under 65 y/o to undergo DXA scans. The desired outcome is to provide the benefits of osteoporosis treatment when appropriate and avoid unnecessary expense, stress, and radiation for women not at significant risk. EHRs are utilized to identify the two populations.

**The evidence relates directly to this process measure. The rationale is to reduce the number of unnecessary DXA scans by requiring use of a risk assessment tool before ordering bone measurement.

**good evidence base supporting measure

**The evidence directly raes to the specific measure process.

**There is a direct relationship.

**The evidence supporting this measure is moderate.

1b. Performance Gap

Comments:

**Yes performance data present and it demonstrated a gap of care with opportunity of improvement and variation. Race/Ethnicity + Age also showed opportunity for improvement and variation.

**They cited both variation in practice, as well as significant rates of overuse. There were significant racial disparities.

**Overuse (18.4% in women 55-59, ~25% in women 60-63); overuse greater in caucasian and asian women.

**There is a gap demonstrated. Disparities among groups are also identified.

**Performance gap with disparities has been demonstrated.

**Performance gap and disparities exist. Performance measure will be useful to attempt to try reducing performance and disparities gaps.

**Data submitted from the literature and from studies at three health care institutions indicate DXA scans are overused in the target population, but at low rates. Both the studies and the literature indicate differences in the steps of care between population subgroups.

**Current performance data was provided. There is overall less than optimal performance for use of clinical osteoporosis risk assessment tools for determine appropriate referral for DXA scans. A national performance measure would provide the focus needed to improve appropriate referrals for bone measurement testing. Data on the measure by population subgroups was provided and showed that race and ethnicity factored in referrals and treatment.

**Current performance data was provided: 6.7% potentially unnecessary DXA scans done in women 50 to 64 y in a large health care system (over 7 million women); another study potentially 40% of DXA scans done in women who do not frailty criteria; finally a retrospective study done at three sites showing potentially inappropriate DXA scans in women between ages 18 and 64. This also showed a gap in care higher rates of overuse in Asian and white women. Also disparities in care noted in that treatment of osteoporosis is lower in

African Americans and Latinos. A national performance measure would focus attention on proper use of DXA and awareness of the contributing risk factors.
**gap identified: 18% - 24% overuse varying by age group; racial disparities
**There is a performance gap based on age and race demonstrating potential disparities in the care being delivered.
**Current performance data is provided. It demonstrates a substantial gap in care and among women of differences races / ethnic groups.
**There is evidence of overuse of DXA scans, providing justification for this measure

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

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

2c. For composite measures: empirical analysis support composite approach

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.

eCQM Technical Advisor review:

Submitted measure is an HQMF compliant eCQM	The submitted eCQM specifications follow the industry accepted format for eCQM (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF,QDM, or CQL limitations	N/A – All components in the measure logic of the submitted eCQM are represented using the HQMF, QDM, or CQL standards
Value Sets	The submitted eCQM specifications uses existing value sets when possible and uses new value sets that have been vetted through the Value Set Authority Center (VSAC).
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously. – this includes 100% coverage of measured patient population testing with pass/fail test cases for each population

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

Evaluators: Staff

Evaluation of Reliability and Validity:

- Score level reliability testing was conducted, and the results indicate that measure is reliable for clinicians with at least 20 patients in the denominator.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?

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

Evaluation A: Scientific Acceptability

Measure Number: 3475e

Measure Title: Appropriate Use of DXA Scans in Women Under 65 Years Who Do Not Meet the Risk Factor Profile for Osteoporotic Fracture

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.**

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 VALIDITY testing** of patient-level data conducted?

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

   Submission document: Testing attachment, section 2a2.2

   - Random split-half correlation, supplemented with bootstrapping. Done at LOA; three sites, approximately 126,000 women; also tested with claims data for 7.5 million women from one large health plan. Developer notes that there are potential issues with the sample used for testing:

     "We pursued testing sites that captured data elements for the measure in their existing EHR workflows. As a result, we recruited sites that could be considered advanced EHR users, suggesting that they are unlikely to be representative of the broader field of clinicians who treat the population of interest. Our approach thus offers evidence that the measure concept is achievable but does not provide conclusive evidence regarding the ability of all EHR users to implement these measures."
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - Average reliability coefficient, for providers with at least 20 patients in the appropriate age range: 0.82. Indicates that measure is reliable for clinicians with at least 20 patients in the denominator.
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.**

    Score level testing with large sample showed reliability.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

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

    Submission document: Testing attachment, section 2b4.

    None

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.

Distribution of missing data not explicitly tested but developer describes how the level of missing data can be found. Developer states:

> In the data files submitted by test sites, there was no distinction between a negative (for example, confirmation that the patient was diagnosed with osteoporosis) and missing data. Where sites reported data for at least one patient, we assumed that blank records indicated no relevant data for those patients. For example, we assumed a patient with no data indicating osteoporosis did not have osteoporosis; we did not exclude that patient from the denominator based on lack of data regarding osteoporosis.

- Kappa could not be calculated for all data elements at all sites due to low prevalence of many exclusions; however, they contribute to the measure's face 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**

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

    Developer drew a random sample and extracted data elements, and then compared with data manually abstracted data. Developer then assessed validity using kappa agreement. Results were stratified by site.
20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    Chance-adjusted agreement between the EHR and manually abstracted data for the numerator was high at two sites (0.91 and 0.93) and extremely low at one site (-0.01: chart prevalence of 48.5% and EHR prevalence of 0.5%). The developer states this is "attributable to a lack of EHR documentation for DXA scans in structured fields"; however the feasibility scorecard, updated more recently, does indicate most data are available in structured fields in EHRs.

    The developer states that chance-adjusted agreement for the denominator exclusions was not reliable due to low prevalence.

    Staff concern: A lack of EHR documentation at one of the three testing sites raises concerns with the measure's validity and feasibility.
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 are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

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

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.**

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

2a1. Reliability-Specification

Comments:

**data element well defined along with codes-descriptors are provided. SDOH may impact ability to reach minorities--they addressed their concern as well.
**Most practices would have 20 patients in the denominator, which is the threshold for reliability. This is reasonable for implementation.
**Missing data
**Data elements are clearly defined. This measure could be consistently implemented.
**No identified concerns.
**No major concerns. Specifications are appropriate
**Reliability specificiations are clear. However, the experience at the four testing sites in which all sites could not complete data analysis with data already included and extratable from their EHRs raises concerns about consistent implementation
**According to the NQF staff review the data elements are clearly defined. The other specifications are clear. If the data can be put into the EHR it is likely to be consistently implemented.
**I have no concerns
**good reliability however, reliability was tested at what was considered "high level" EHR users which may not be representative of typical EHR users;
**The specifications are clear and should be able to be consistently implemented for measurement purposes.
**No concerns
**No concerns

2a2. Reliability-Testing

Comments:

**Not at this time
**No
**No
**None
**No concerns.
**No major concerns. Reliability is high
**See comment on 6.2a1.
**the developer indicated that test sites were chosen where the clinicians were thought to be expert users of an EHR. This indicated to them that the measure could be reliably implemented at least at these sites with advanced EHR users, but that it does not necessarily mean that all EHR users could implement the measure

**NQF staff indicated that reliability testing met if clinic had at least 20 patients in the denominator, therefore I have no concerns about this
**none
**No concerns with the reliability testing.
**It is not clear that all EHR practices will have access the needed data elements.
**No concerns

2b1. Validity-Testing

Comments:

**No
**No
**Missing data? and its impact.
**None.
**No concerns.
**Assumptions regarding lack of data-meaning lack of exclusions-may be problematic
**No
**Chance adjusted agreement for the numerator was performed at three sites correlating EHR data with manually abstracted data. Two of the three clinics had high correlations, one was very poor. The developers indicated that the problem was due to inadequate EHR support for entry of DXA information. Validity testing for the denominator compromised by low prevalence of the exclusions.
**The developers looked at random samples and correlate data abstracted from paper records and EHR data. Two of the three sites had high correlation; the third site had low correlation. Developers attributed that to lack of DXA documentation facility on the existing EHR. If the developer is correct, this issue will fade as most clinicians/groups use robust EHR's
**Developer makes assumption that if nothing in the record regarding osteoporosis then patient is assumed to not have risk factors and is therefore not excluded from denominator; perhaps review some of these cases specifically (by manual review) to confirm that they should not be excluded from denominator;
**No concerns with the validity testing.
**No
**No concerns

2b4-7. Threats to Validity

2b.4. Meaningful Differences

Comments:

**Not at this time
**The concern over missing data due to EHR is a concern.While this is not common, it can not be assumed it could not happen at other sites.
**perhaps as interpreted by the developer (no data interpreted as no osteoporosis)
**No majors threats.
**No recognized threats to validity.
**yes, this can be assumed to be a threat unless there's information to suggest otherwise.
**Missing data: It is assumed that a patient with no data indicating osteoporosis did not have osteoporosis. This is consistent with the current state of medical records but is not necessarily true.
**Higher scores will indicate higher potentially inappropriate referrals. 2b5 There is only one set of specifications. 2b6 Missing data is treated as lack of an osteoporosis diagnosis, or lack of adequate score needed to reliably order a DXA
**Missing data will be treated as "no osteoporosis". There is wide variation among clinicians/clinics in the use of DXA in women who do not meet the requirement for use of a formal clinical risk assessment tool. There is a large number of potential exclusions and for any one clinic the number of individuals who represent those exclusions may be low.
**see above
**I did not identify any threats to the validity of the measure or to measure results.

**2b.6 There was no distinction between a negative and missing data which may constitute a threat to validity.A lack of EHR documentation at one of the three testing sites raises concerns with the measure's validity and feasibility.
**No concerns

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment

Comments:

**Exclusions are consistent with evidence.
**n/a
**Deemed no applicable by developer
**Exclusions are appropriate.
**No recognized other threats to validity.
**No other issues or concerns
**Exclusions appear to be consistent with the current evidence and complete. No risk adjustment or stratification.
**2b2 No groups are excluded. 2b3 there is a conceptual relationship between potential social risk factor variables and the measure focus. Potentially inappropriate DXA scans are performed more frequently in Asian and white women than in African American and Latino women.
**Exclusions are consistent with evidence. There are some women who lose bone mineral rapidly at menopause who do not have any of the accepted risk factors. There is no way to capture this population at present without bone measurement. There is a clear conceptual relationship between potential social risk factor variables and the measure focus. The risk-adjustment variables were present at the start of care. Risk adjustment was properly developed and tested. Results are acceptable and there is an appropriate risk-adjustment strategy included in the measure.
**none
**There appeared to be some challenges with abstacting data correctly from the EHR and some conflicting information on whether that was due to the EHR field structure or how information was documented in the EHR.
**Certain exclusions (gastric bypass for example) could not be completely evaluated.
**No concerns

Criterion 3. Feasibility

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

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

- The data for this measure are generated/collected by healthcare providers during the provision of care.
- All data elements are available in EHRs.
- The measure is an eMeasure and has been reviewed by the eMeasure team.
- The FRAX can be accessed online for free or a desktop edition can be purchased. The measure is available for public use and there are no other fees associated.

Feasibility Testing	Number of data elements included in measure calculation: 41 Number of data elements scoring less than 3 on scorecard: 31 **Questions for the Committee:** Consider the following questions for each data element that scored less than a 3 in any of the feasibility domains: *How is the data element used in computation of measure?* *How the data element is feasible within the context of the measure logic?* *What is the plan for readdressing the data element?* The following data elements scored less than 3 in the Workflow domain at one of the four sites: • Diagnosis: Type 1 Diabetes • Diagnosis: Rheumatoid Arthritis • Diagnosis: Psoriatic Arthritis • Diagnosis: Ankylosing Spondylitis • Diagnosis: Ehlers Danlos Syndrome • Diagnosis: Marfan's Syndrome • Diagnosis: Osteopenia • Diagnosis: Osteoporosis • Diagnosis: Osteogenesis Imperfecta • Diagnosis: Osteoporotic Fractures • Diagnosis: Cushings Syndrome • Diagnosis: Lupus • Diagnosis: Hyperthyroidism • Diagnosis: Hyperparathyroidism • Diagnosis: Chronic Liver Disease • Diagnosis: Chronic Malnutrition • Diagnosis: Malabsorption Syndromes • Diagnosis: End Stage Renal Disease Data Element 1 • List low scoring domains: Availability – Accuracy – Standards - Workflow How is the data element used in computation of measure? How the data element is feasible within the context of the measure logic? What is the plan for readdressing the data element? The following data elements scored less than a 3 in the Availability, Accuracy, Standards, and Workflow domains in at least of two of the sites: • Risk Category Assessment: History of hip fracture in parent • Feedback from developer's Feasibility Assessment: *Clinicians at three sites did not collect the history of hip fracture in a parent in a structured field in the EHR. Two of these sites did not have a structured field for this element, and they did not consistently inquire about it as part of the clinical workflow. The other site could capture this element in a structured field, but clinicians did not always ask about it as part of the patient's medical history. Therefore, the record was not always accurate and would require a workflow change to ensure routine documentation of the data element by all providers.* • Risk Category Assessment: Ten-year probability of all major osteoporosis related fracture (FRAX Score)

	• Feedback from developer's Feasibility Assessment: *One of the four sites used FRAX to determine whether to order a DXA scan, but clinicians at this site did not document the FRAX score in a structured field in the EHR. Test-site staff noted that if clinicians were to start documenting this score in the EHR, it would most likely be entered as free text.* *Clinicians at two other sites used the FRAX tool—but not to determine when to order DXA scans. Instead, they typically ordered DXA scans and calculated the FRAX score afterward, using information from the scans (such as bone-mineral density) as an input. Clinicians cited two reasons for using the FRAX tool after receiving the scan results: (1) some believe that bone-mineral density was a required input to calculate the FRAX score, although it is actually optional, and (2) clinicians felt that the FRAX score would be more accurate if the DXA scan results were included. At one of these sites, clinicians typically entered the score as free text in the EHR, which was linked to a diagnosis (such as osteoporosis) and a date and time. Clinicians at the other site entered the score in a structured field.* *Staff at the fourth site, which was not using the FRAX tool, said that they hoped the tool would be incorporated into the clinical workflow and EHR in the next one or two years, but they noted that the scores would also most likely be entered post-DXA scan to determine the appropriate treatment for patients.* The following data element scored less than 3 in the Data Accuracy and Workflow domains in 2/4 sites: • Risk Category Assessment: Average Number of Drinks per Drinking Day • Feedback from developer's Feasibility Assessment: *Clinicians at three sites administered the Alcohol Use Disorders Identification Test (AUDIT) as a screening questionnaire for alcohol abuse in patients, and this questionnaire includes a question about the average number of drinks per drinking day. Clinicians at one of these sites only recently started administering the AUDIT for new patients and estimated that the results were available in their EHR only for 25 percent of patients. Clinicians at the fourth site documented the average drinks per day, but not per drinking day, in a structured field in the EHR. As with all measures that require self-reported information on substance use, data accuracy is an issue at all of these sites because patients might not provide truthful answers about their use. However, where AUDIT is consistently used and the results are stored in structured fields, the data element is available and feasible to extract.* The following data element scored less than 3 in the Data Accuracy domain in 3/4 sites: • Medication, Active: Aromatase Inhibitors • Medication, Order: Aromatase Inhibitors • Medication, Active: Glucocorticoids (oral only) • Medication Dosage, Glucocorticoids (oral only) • Medication Duration: Glucocorticoids (oral only) Feedback from developer's Feasibility Assessment: *The measure excludes patients who have taken aromatase inhibitors at any point during their history and considers whether a patient has taken 5 mg per day or more of oral glucocorticoids over a period of at least 90 days at any point during their history. Although sites captured active medications and medication orders in structured fields in their EHRs, test-site staff said that medication reconciliation does not always occur. Therefore, the EHR might not accurately reflect when patients stop taking medications. Medication history for new patients or patients seen by external providers might also be incomplete in the EHR.* *Clinicians at one site could find medications that had been e-prescribed or could manually enter new patients' medication history from transferred medical records. At another site, clinicians could request a list of medications for the patient from the pharmacy, but only for the past two*

	years. A third site switched to a new EHR in January 2017, and site staff said that previous medical information was transferred inconsistently, resulting in an incomplete medication history for some patients. *In addition, none of the sites captured the daily dosage of active or ordered medications in a structured field, but providers routinely documented prescription quantity, strength (for example, 5 mg per pill), and number of refills in structured fields. The frequency of medications (for example, two pills a day) was documented as free text at three sites. Because not all inputs necessary to calculate daily dosage are available in structured fields of the EHR, manual calculation would be required to determine the daily dosage for oral glucocorticoids, and these calculations would be subject to error. Furthermore, two sites did not have structured fields for the stop and start dates of medications; practices would therefore need to calculate the duration of active and ordered medications based on refill dates, which could reduce accuracy.* The following data element scored less than a 3 in the Availability, Accuracy, Standards, and Workflow domains at two sites: • Procedure, Performed: Gastric Bypass Surgery Feedback from developer's Feasibility Assessment: *Two of the four sites captured gastric bypass surgeries in structured fields of the EHR and indicated that it was feasible to use this data element to exclude patients from the measure. The other two sites, both using GE Centricity, did not consistently capture gastric bypass surgery in structured fields. Clinicians at one of the sites documented gastric bypass surgery as free text, and staff at the other site said that clinicians did not always ask about gastric bypass surgery, so documentation depended on whether the patient volunteered the information or if the clinician was involved in the patient's care at the time of the procedure.* The following data elements scored less than a 3 in the Data Standards domain at one site: • Encounter, Performed: Face-to-Face Interaction • Patient Characteristic Race: Race • Patient Characteristic Payer: Payer • Patient Characteristic Ethnicity: Ethnicity

Questions for the Committee:

- Does the Committee think the identified feasibility issues are fixable, as suggested by the developer, or do they raise larger concerns around the measure's overall feasibility?
- Is it reasonable to assume providers will be able to modify EHRs and/or clinical workflows to accurately report the measure?
- Do the developer's plan for the issues encountered in testing suffice?
- Is the data collection strategy ready to be put into operational use?
- Does the eCQM Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?

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

RATIONALE:

- This measure scored low on feasibility due to a number of potential issues. NQF staff have identified potential issues the Committee should discuss.

- The measure received mixed results at the four testing sites. None of the sites were able to fully implement the measure. However, the developer notes this is a new measure and that minor changes to workflow or products should allow sites to capture all of the data elements needed.
- There are 41 data elements included in the measure, of which 31 scored less than three in at least one of the four testing sites.
- NQF's eMeasure Feasbility Report states that if any data element scores as 1, the data element has low feasibility, regardless of summary scores. Three of the data elements scored a 1 at all four test sites.

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility

Comments:

**well documented and tested

**not all reports come back in form of structured data, so could require manual entry. frax tool is not regularly documentee in emr, so could create extra steps.

**Screening tools to assess risk are, in many practices not part of the EHR. Using risk calculators can provide estimate of patient risk but may not be found in a common data element in the EMR, thus leading to challenges with having the risk score that purportedly may have initiated the test.

**Barriers identified during feasibility assessment are FRAX calculation, Gastric bypass surgery, parental hip fx history. Largely due to not in a discrete field in the ehr.some variability due to type of ehr used.

**No identified concerns about feasibility.

**The identified feasiblity issues may be problematic. Additional discussion during committee call is needed. Providers are able to modify EHRs and clinical workflows to accurately report the measure; however, whether they WILL is another issue entirely. The developers plan to counter the issues seem sufficient, but further discussion during committee call is needed.

**Feasibility is worrisome as many of the required data elements (including diagnoses leading to exclusions & current medications lists were not available to be extracted by the EHR. This is proposed as a MIMS addition. I suppose if this proposed measure does not require much more "manual" extraction of data by reading the record for free text entries, etc., than current MIPS measures, then this problem might not be a "deal breaker".

**The required data elements are routinely generated during care delivery. If the clinic has a robust EHR, then the required data elements should be available. This measure can be put into operational use when the proper EHR software installed

**5 exclusion criteria (FRAX, hip fracture in parent, gastric bypass, medication reconciliation, # drinks/day) were difficult to capture and could influence results; the developer needs to reconcile how these 5 elements could be captured in spite of EHR challenges;

**There was significant inconsistency across testing sites indicating challenges with the feasibiity of collecting the data necessary to calculate the measure.

**I'm concerned about inconsistent collection of the history of hip fracture in a parent in a structured field in the EHR. Also, inconsistent use of FRAX tool, that medication reconciliation does not always occur, and thatnone of the sites captured the daily dosage of active or ordered medications in a structured field.

**Concerns were raised over feasibility, but these concerns seem to be addressable

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details

This measure will be in MIPS starting in 2019.

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

Feedback on the measure by those being measured or others

The measure is not yet in use so hasn't been shared with anyone being measured. However, results were shared with test sites during testing as well as with the technical expert panel and a DXA Overuse expert work group. None of these groups had any significant concerns about their performance/clinician performance on the measure.

Additional Feedback:

Not available

Questions for the Committee:

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

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

RATIONALE:

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

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

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

Improvement results

The measure is not yet in use so no improvement results were submitted.

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

This measure has not yet been implemented so no results are available, but the developers note that it should encourage providers to follow the USPSTF guideline because it will encourage the use of clinical risk assessment tools and because it may "increase clinicians' consistency in determining which patients are at high risk for osteoporotic fracture—and therefore eligible for a DXA scan."

Potential harms

Potentially, the measure could cause women who do not have the risk factors identified, or not enough of them, to miss needed DXA scans and therefore not receive or be delayed in receiving needed treatment. Also, the screening tool (FRAX) has not yet been widely studied in nonwhite groups, so women of color could not receive appropriate treatment, or have delays in receiving treatment.

Additional Feedback:

Not available

***Questions for the Committee**:*

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

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

RATIONALE:

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

4a.1 Use-Accountability and Transparency

Comments:

**not public; feedback from the technical expert panel and a DXA Overuse expert work group--one of these groups had any significant concerns

**unsure

**Does not appear to be publicly reported nor does there appear to be opportunities for feedback

**Credible plan. Feedback not yet given as not in use but centers in feasibility given feedback and concerns not voiced .

**I'm not aware of public reporting or feedback to those being measured.

**It is an ambitious measure, with feasibility problems. The issues related to "USE" can be addressed only if the solutions to address feasibility are implemented, and if they work.

**Credible plan for implementation. "The two test sites did not share any significant concerns about their performance on their measure." Were providers and staff asked to comment on the process of the measuring?

**There are 41 data elements required for this measure. There were 31 that scored less than 3 in the workflow domain of 1 of 4 clinics. About 19 data elements in particular scored less than 3 at one clinic. Workflow change will have to occur at 2 of the 4 clinics to accurately indicate hip fracture in a parent. All four of the clinics had some issue, actually different issues employing the FRAX tool at present. Alcohol use (drinks per day) and glucocorticoid and aromatase inhibitor use could not always be reliably assessed from review of the EHR. Gastric bypass surgery data capture was inadequate. NQF staff commented LOW feasibility. Developer thinks all of this can be fixed with EHR upgrade.

**not publicly reported yet;
**The measure is being used in an accountability program and will be included in MIPS in 2019.
**No concerns.
**No concerns.

4b1. Usability-Improvement

Comments:

**I agree with their statement: "increase clinicians' consistency in determining which patients are at high risk for osteoporotic fracture—and therefore eligible for a DXA scan."
**would require working with emr vendor to have better way to document use of frax tool
**The measure has importance for over utilization of testing, over utilization of treatment, with potential for long term medication use that can be expensive with little overall patient benefit. Mitigation of individual harm and improving healthcare value are important potential benefits.
**Reducing unnecessary testing would contribute to high-quality case. Only potential harm is reduction in DXA use among in patient appropriate for the test.
**I am not aware of actual unintended consequences.
**Uninted harms are appropriately listed. No additional comments from me.
**The rationale of how results would further healthcare improvement seems solid. As described, benefits outweigh harms.
**If this measure can be successfully used DXA overuse will be reduced. Hopefully more individuals who would benefit from getting the test and subsequent treatment will be served. The unintended consequence is that women who experience rapid and severe bone mineral loss after menopause who do not meet the requirement for score/age or special disease consideration will be missed. Overall, considering that many DXA are currently incorrectly performed and reported and that inappropriate treatment may result, the benefits outweigh the harms.
**not publicly reported yet; there is need (not necessarily responsibility of the developer) to more carefully study use of FRAX across different populations (race disparity)
**The results of the measure could reduce unnecessary testing, inappropriate or unneeded treatment. May also create cost-efficiencies in treating osteoporosis.
**My only concern is the measure could cause women who do not have the risk factors identified, or not enough of them, to miss needed DXA scans and therefore not receive or be delayed in receiving needed treatment.
**No concerns.

Criterion 5: Related and Competing Measures

Related or competing measures

This measure is related to 0046 Screening or Therapy for Osteoporosis for Women Aged 65 Years and Older: Percentage of female patients aged 65-85 years of age who ever had a central dual-energy X-ray absorptiometry (DXA) to check for osteoporosis.
The developer states that the two measures complement each other.

Harmonization

The measures are harmonized to the extent possible, but have significant differences:

- The measures are different levels of analysis: 0046 is for claims and registry LOA; this measure is clinician LOA
- The measures have different intents: 0046 assesses documentation of DXA results, and is limited to DXA scans of the hip or spine (central DXA scans); 3475e assesses DXA orders for both central and peripheral scans
- The measures cover different populations: 0046 is for women ages 65 and older; 3475e is for women under 65

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

5. Related and Competing

Comments:

**I agree with their statement: This measure is related to 0046 Screening or Therapy for Osteoporosis for Women Aged 65 Years and Older: Percentage of female patients aged 65-85 years of age who ever had a central dual-energy X-ray absorptiometry (DXA) to check for osteoporosis. The developer states that the two measures complement each other.

**No

**no need for harmonization

**The other measure regarding DXA screening in patients 65-85 is in a different age group so doesn't compete or harmonize.

**I am not aware of any competing measures.

**Complements 0046 well

**NQF 0046 is different but related and not competing.

**#3475e is related to #0046 screening or rx for Osteoporosis in women >= 65: % females 65-85 who ever had a central DXA to check for Osteoporosis. The developer states these measures have been harmonized to the extent possible but they differ in level of analysis, intents, and population age. No additional harmonization is possible without major changes in the measures

**no concerns

**I am not aware of competing or related measures.

**Measures appear complementary.

**No concerns

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January/25/2019
No NQF members who have submitted a support/non-support choice

Brief Measure Information

NQF #: 3475e

Corresponding Measures:

De.2. Measure Title: Appropriate Use of DXA Scans in Women Under 65 Years Who Do Not Meet the Risk Factor Profile for Osteoporotic Fracture

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services, Center for Clinical Standards and Quality, Quality Measurement and Value-Based Incentives Group (QMVIG), Division of Electronic and Clinician Quality, MS S3-02-01

De.3. Brief Description of Measure: Percentage of female patients 50 to 64 years of age without select risk factors for osteoporotic fracture who received an order for a dual-energy x-ray absorptiometry (DXA) scan during the measurement period.

1b.1. Developer Rationale: This measure is expected to increase the recording of patient risk for fracture data and to decrease the number of inappropriate DXA scans. Current osteoporosis guidelines recommend using bone measurement testing to assess osteoporosis risk in women ages 65 and older. In postmenopausal women younger than 65, guidelines recommend using a formal clinical risk assessment tool to establish patients' risk for osteoporosis in order to determine whether to screen them for osteoporosis using bone measurement testing. Clinical information such as age, BMI, parental history of hip fracture, smoking, and alcohol use can be used to determine a woman's fracture risk (U.S. Preventive Services Task Force, 2018).

In addition, there are potentially avoidable harms associated with screening for osteoporosis in general, including exposure to radiation, false-positive exams, and the side effects of unnecessary osteoporosis medications, which add costs to an already burdened health care system (Lim et al., 2009).

Citations:

Lim LS, Hoeksema LJ, Sherin K. Screening for osteoporosis in the adult U.S. population: ACPM position statement on preventive practice. Am J Prev Med. 2009;36(4):366-75.

U.S. Preventive Services Task Force. Screening for osteoporosis to prevent fractures: U.S. Preventive Services Task Force recommendation statement." JAMA. 2018;319(24):2521-31.

S.4. Numerator Statement: Female patients who received an order for at least one DXA scan in the measurement period.

S.6. Denominator Statement: Female patients ages 50 to 64 years with an encounter during the measurement period.

S.8. Denominator Exclusions: The measure excludes patients who have a combination of risk factors (as determined by age) or one of the independent risk factors.

De.1. Measure Type: Process: Appropriate Use

S.17. Data Source: Electronic Health Records

S.20. Level of Analysis: Clinician : Individual

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. This measure is not paired or grouped.

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

DXA_Evidence_Attachment_Final-636772656013050280.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*)**:**

Measure Title: **Appropriate Use of DXA Scans in Women Under 65 Years Who Do Not Meet the Risk Factor Profile for Osteoporotic Fracture**

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

Date of Submission: 11/8/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: 3 Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: 5 a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured structure leads to a desired health outcome.
- Efficiency: 6 evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☐ Outcome:

☐Patient-reported outcome (PRO):

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

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

☒ Process:

☒ Appropriate use measure: Overuse of Dual-Energy X-Ray Absorptiometry (DXA) Scans in Women Under 65 Who Do Not Have Select Risk Factors for Osteoporotic Fracture

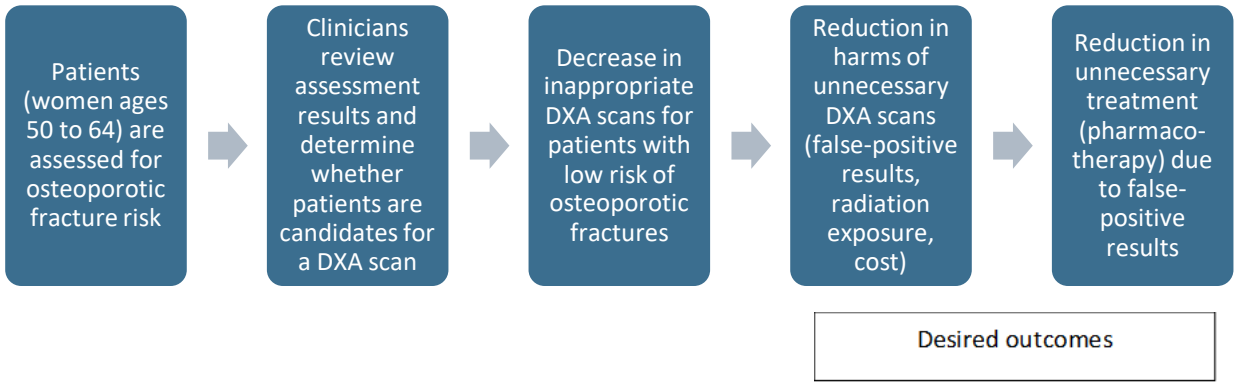
☐ Structure:

☐ Composite:

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

The goals of this overuse measure are to (1) decrease inappropriate DXA screenings for osteoporosis and (2) reduce avoidable harms associated with screening patients who have a low risk of osteoporotic fractures.

The presumed pathway from process to outcomes is as follows:



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

Not applicable. This measure does not rely on patient-reported data.

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

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

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	U.S. Preventive Services Task Force (USPSTF) recommendation: • Osteoporosis to Prevent Fractures: Screening • USPSTF • June 2018 • USPSTF. Screening for osteoporosis to prevent fractures: U.S. Preventive Services Task Force recommendation statement. JAMA. 2018;319(24):2521-31. • https://jamanetwork.com/journals/jama/fullarticle/2685995 Evidence review supporting USPSTF recommendation: • Screening to Prevent Osteoporotic Fractures: Updated Evidence Report and Systematic Review for the U.S. Preventive Services Task Force • Viswanathan, M., Reddy, S., Berkman, N., Cullen, K., Middleton, J., Nicholson, W., and Kahwati, L. • June 2018 • Viswanathan M, Reddy S, Berkman N, Cullen K, Middleton J, Nicholson W, et al. Screening to prevent osteoporotic fractures: updated evidence report and systematic review for the U.S. Preventive Services Task Force. JAMA. 2018;319(24):2532-51. • https://jamanetwork.com/journals/jama/fullarticle/2685994
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.	"The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in postmenopausal women younger than 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool." (USPSTF, 2018)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The USPSTF does not grade the evidence. They review the evidence identified through the evidence review and determine if the benefits outweigh the harms. For the recommendation grade, see "Grade assigned to the recommendation with definition of the grade" below.
Provide all other grades and definitions from the evidence grading system	The USPSTF does not grade the evidence. They review the evidence identified through the evidence review and determine if the benefits outweigh the harms. For the grading system used by the USPSTF, see "Provide all other grades and definitions from the recommendation grading system" below.
Grade assigned to the **recommendation** with definition of the grade	"The USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in postmenopausal women younger than 65 years who are at increased risk of osteoporosis is at least moderate." The USPSTF recommendation is a grade B recommendation. Grade B—There is high certainty that the net benefit is moderate, or there is moderate certainty that the net benefit is moderate to substantial.

Provide all other grades and definitions from the recommendation grading system	The USPSTF used the following system for grading the body of evidence: • Grade A—The USPSTF recommends the service. There is high certainty that the net benefit is substantial. • Grade B—Grade B is described above. • Grade C—The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. • Grade D—The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. • I statement—The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	In 2018, Viswanathan et al. (2018) conducted a systematic review to support the USPSTF as it considered an update to its 2011 recommendation for osteoporosis screening. Viswanathan and colleagues reviewed the evidence published from November 2009 to October 2016 to identify evidence published since the 2011 review on screening for and treatment of osteoporotic fractures, risk assessment tools, and the efficacy of screening. Unless otherwise noted, we obtained information on the quality and quantity of the studies from this evidence review. Overall, the evidence review captured 168 published articles of good or fair quality. The USPSTF uses the following criteria to rate the quality of the evidence: "Randomized controlled trials and cohort studies • Initial assembly of comparable groups: o For randomized controlled trials: adequate randomization, including first concealment and whether potential confounders were distributed equally among groups o For cohort studies: consideration of potential confounders, with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts • Maintenance of comparable groups (includes attrition, cross-overs, adherence, contamination) • Important differential loss to follow-up or overall high loss to follow-up • Measurements: equal, reliable, and valid (includes masking of outcome assessment) • Clear definition of interventions • All important outcomes considered • Analysis: adjustment for potential confounders for cohort studies or intention-to-treat analysis for randomized controlled trials "Definitions of ratings based on [the] above criteria: Good: Meets all criteria: Comparable groups are assembled initially and maintained throughout the study (follow-up ≥80%); reliable and valid measurement instruments are used and applied equally to all groups; interventions are spelled out clearly; all important outcomes are considered; and appropriate attention [is paid] to confounders in [the] analysis. In addition, intention-to-treat analysis is used for randomized controlled trials. "Fair: Studies are graded 'fair' if any or all of the following problems occur, without the fatal flaws noted in the 'poor' category below: Generally comparable groups are assembled initially, but some question remains [about] whether some (although not major) differences occurred with follow-up; measurement instruments are acceptable (although not the best) and generally applied equally; some but not all important outcomes are considered; and some but not all potential confounders are accounted for. Intention-to-treat analysis is used for randomized controlled trials. "Poor: Studies are graded 'poor' if any of the following fatal flaws exists: Groups assembled initially are not close to being comparable or maintained throughout the study; unreliable or invalid measurement

	instruments are used or not applied equally among groups (including not masking outcome assessment); and key confounders are given little or no attention. Intention-to-treat analysis is lacking for randomized controlled trials." (Viswanathan et al., 2018). The information on evidence quantity and quality is organized by key questions assessed in the evidence review. *Key question 1. Does screening (clinical risk assessment, bone density measurement, or both) for osteoporotic fracture risk reduce fractures and fracture-related morbidity and mortality in adults?* The authors identified one fair-quality controlled study that evaluated the effect of screening for hip-fracture risk and treatment. *Key question 2a. What is the accuracy and reliability of screening approaches to identify adults who are at increased risk for osteoporotic fracture?* Clinical risk assessment tools, like the FRAX, can be used to identify osteoporosis or to assess a person's risk for osteoporotic fracture. The authors reviewed the evidence for both uses of clinical risk assessment tools. However, this measure focuses on the use of the FRAX to assess fracture risk, and thus we present the review of the evidence specific to this use of risk assessment tools. The authors included one good-quality systematic review that assessed the accuracy of clinical risk assessment tools in predicting fracture in adults. This systematic review included 45 articles that assessed 13 risk-prediction tools. Twenty-six studies assessed the FRAX, six assessed the Garvan Fracture Risk Calculator, and four assessed the QFracture prediction tool. Other tools were assessed by only one or two studies each. The authors also identified and included in the evidence review 13 observational studies with low risk of bias or unclear bias that were not included in the systematic review, either because they were published after the systematic review search dates or because they were not identified or included in the systematic review. *Key question 3: What are the harms of screening for osteoporotic fracture risk?* The single fair-quality controlled study identified to answer this question is the same study referenced above in key question 1. *Key question 5: What are the harms associated with pharmacotherapy?* One of the potential harms from the overuse of a screening test is the downstream effects for patients who have a positive screening result. For DXA screening, a positive test (an osteoporosis diagnosis) could lead to the use of pharmacotherapy. The authors identified 16 fair- and good-quality studies reporting on the harms of alendronate, 4 fair- and good-quality studies on zoledronic acid, 6 fair-quality studies on risedronate, 2 fair-quality studies on etidronate, 7 fair-quality studies on ibandronate, 6 good-quality studies on raloxifene, 4 fair-quality studies on denosumab, and 1 fair-quality study on parathyroid hormone.

Estimates of benefit and consistency across studies	The USPSTF concluded "with moderate certainty that the net benefit of screening for osteoporosis in postmenopausal women younger than 65 years who are at increased risk of osteoporosis is at least moderate." This was partly based on the evidence for key questions 1, 2a, and 5. We provide the evidence reviews for these key questions below. The USPSTF does not have a specific recommendation on the overuse of DXA (that is, it does not explicitly state when not to screen women for osteoporosis). In the evidence review supporting the USPSTF recommendation, the authors assessed the evidence for the harms associated with osteoporosis screening (key question 3). They found one fair-quality controlled study that evaluated how screening for hip-fracture risk and treating those at high risk affects fracture rates in postmenopausal women ages 70 to 85. Participants in the intervention group were initially assessed for 10-year hip-fracture risk using the FRAX, and if the FRAX identified them as high risk, they were offered a DXA screening. Women were then offered treatment, as appropriate, based on the results of the DXA test and a revised FRAX (which incorporated the DXA results). This study showed no differences in anxiety or quality of life between participants in the intervention group versus the control group. However, the USPSTF notes that the potential harms of screening for osteoporosis include "false-positive test results, which can lead to unnecessary treatment, and false-negative test results" as well as "radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)." Although the USPSTF did not specifically recommend against the use of DXA screening for osteoporosis in women at low risk for osteoporotic fracture, it did recommend osteoporosis screening (using bone measurement testing) only in postmenopausal women younger than 65 who are at increased risk of osteoporosis, "as determined by a formal clinical risk assessment tool." This measure attempts to identify women who are not at increased risk for osteoporotic fracture and assesses whether they were potentially inappropriately screened for osteoporosis using a DXA scan. One exclusion for the measure is a FRAX score indicating a high risk of osteoporotic fracture. *Key question 1. Does screening (clinical risk assessment, bone density measurement, or both) for osteoporotic fracture risk reduce fractures and fracture-related morbidity and mortality in adults?* The one study identified for this question is the same as the study identified for key question 3 (described above). According to the evidence review, "this study reported no significant difference in the primary outcome of any osteoporotic fracture in women screened with FRAX compared to women receiving usual care." In addition, the study did not show a statistically significant difference for all clinical fractures or mortality. However, the study did reveal a statistically significant lower incidence of hip fracture in the intervention group. *Key question 2a. What is the accuracy and reliability of screening approaches to identify adults who are at increased risk for osteoporotic fracture?* Across 12 studies that included 190,795 women, the accuracy of the FRAX (without the use of bone measurement density in the calculation) in predicting hip fractures for women was 0.76. This was similar to or

	higher than the accuracy rates for other clinical risk-prediction tools, which ranged from 0.52 to 0.71 (however, no other tools assess the risk of hip fracture specifically). *Key question 3: What are the harms associated with osteoporosis screening?* The study showed no differences in anxiety or quality of life between participants in the intervention group versus the control group. The USPSTF notes that the potential harms of screening for osteoporosis include "false-positive test results, which can lead to unnecessary treatment, and false-negative test results" as well as "radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)." *Key question 5: What are the harms associated with pharmacotherapy?* • **Bisphosphonates**: "The USPSTF identified 16 studies on alendronate, 4 studies on zoledronic acid, 6 studies on risedronate, 2 studies on etidronate, and 7 studies on ibandronate that reported on harms. Overall, based on pooled analyses, studies on bisphosphonates showed no increased risk of discontinuation, serious adverse events, or upper gastrointestinal events." • **Raloxifene**: "Six trials of raloxifene therapy in women reported on various harms. Pooled analyses showed no increased risk of discontinuation due to adverse events or increased risk of leg cramps. However, analyses found a nonsignificant trend for increased risk of deep vein thrombosis, as well as an increased risk of hot flashes." • **Denosumab**: "Four studies reported on harms of denosumab therapy in postmenopausal women. Pooled analyses showed no significant increase in discontinuation or serious adverse events but found a nonsignificant increase in serious infections." • **Parathyroid hormone**: "A single study of parathyroid hormone therapy in women reported an increased risk of discontinuation and other adverse events, such as nausea and headache" (Viswanathan et al. 2018). In **Section 1a.4. Other Source of Evidence**, we provide information from other sources that demonstrates DXA overuse by clinicians and the unintended consequences of these scans.
What harms were identified?	The USPSTF does not identify significant harms of FRAX assessments or DXA scans. But the task force notes that "potential harms of screening for osteoporosis include false-positive test results, which can lead to unnecessary treatment, and false-negative test results" as well as "radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)." For more information about the harms of unnecessary DXA scans, see Section **1a.4. Other Source of Evidence**.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No additional studies were identified since the publication of the guideline.

1a.4 OTHER SOURCE OF EVIDENCE

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

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

This measure was developed based on the American Academy of Family Physicians' Choosing Wisely recommendation statement on DXA for osteoporosis, which states, "Don't use dual-energy X-ray absorptiometry (DEXA) screening for osteoporosis in women under age 65 or men under 70 with no risk factors" (AAFP, n.d.). This recommendation also encourages clinicians to use a risk assessment tool, such as the FRAX, to determine the need for a DXA scan. Although the recommendation and additional studies described below address overuse of DXA scans in women below age 65, there are no quality measures that assess the overuse of DXA scans.

Evidence in Support of Appropriate DXA Use

Studies suggest that among women who have had a DXA scan, about 40 percent do not meet risk factors for frailty (Schnatz et al., 2011). As a result of the DXA scan, these women may receive inappropriate medication and treatment for osteoporosis or osteopenia. One study showed that up to two-thirds of newly prescribed osteoporosis medications were given based on abnormalities identified using DXA scans that do not meet clinical guidelines for diagnosis (Fenton et al., 2016). (For more information about harms associated with osteoporosis medication, see Section **1a.3. Systematic Review of the Evidence**.) Furthermore, in a study of 451 reports from DXA scans, 80 percent contained an error related to image data analysis (Messina et al., 2015). As patients typically consider bone-scan results to be definitive, this poses problems for overdiagnosis and overtreatment for osteoporosis and osteopenia because patients may not question the findings (Moynihan et al., 2017).

Despite the problems associated with DXA scans, overdiagnosis of osteoporosis and osteopenia, and subsequent inappropriate medication and treatment, clinicians continue to overuse these scans. A retrospective longitudinal analysis conducted across 34 practices showed no difference in the rates of DXA scan usage before and after the publication of the Choosing Wisely recommendation about DXA overuse (Lasser et al., 2016). The authors of this study suggest that "targeted initiatives addressing providers with high ordering rates will be needed to change behavior."

In addition, a retrospective cohort study of 13 practices assessed the three-, five-, and seven-year incidence of inappropriate and appropriate DXA scans. The study team found a three-year incidence of DXA scans of 18.4 percent in women ages 50 to 59 without osteoporosis risk factors, and 24.9 percent in women ages 60 to 64 without risk factors (Amaranth et al., 2015). These studies suggest that a measure targeting appropriate use of DXA scans, as informed by a risk assessment tool, could improve care delivery.

Evidence for Exclusions

This measure includes three types of exclusions: (1) high risk of hip fracture as determined by a FRAX score, (2) conditions or patient characteristics that are used to determine a FRAX score (called "combination" risk factors), and (3) conditions or patient characteristics that are associated with a high rate of osteoporotic fracture. The table below shows risk factors that fall into the third group and that we identified in the literature as having high-risk ratios. Patients with ankylosing spondylitis, for example, have a relative risk of 7.1, which means that this condition is associated with a 700 percent higher chance of an osteoporotic fracture compared with a healthy person's chances.

Exclusion	Risk of fracture
Ankylosing spondylitis	7.1 odds ratio (OR; 95 percent confidence interval [CI]: 6.0–8.4) for vertebral fractures in patients with ankylosing spondylitis (Weiss et al., 2010).
Aromatase inhibitors	In studies comparing the use of aromatase inhibitors versus no aromatase inhibitors, the medication increased fracture risk by 17 percent in women under age 65 (95 percent CI: 1.07–1.28) (Tseng et al., 2018).
Cushing's syndrome	Patients with Cushing's syndrome were significantly more likely to report a low-energy fracture (a fracture occurring after minimal or no trauma) compared with controls (9.5 percent compared with 1.8 percent; $p = 0.004$) (Vestergaard et al., 2002).
Ehlers-Danlos syndrome	Previous fracture was 10 times more common in patients with Ehlers-Danlos syndrome ($p < 0.001$) than in other patients; 86.9 percent of patients with Ehlers-Danlos syndrome reported low-impact fractures (fractures of a peripheral bone) compared with 8.7 percent of controls (Dolan et al., 1998).
End-stage renal disease	4.11 standardized incidence ratio (95 percent CI: 2.96–5.73) for hip fracture in female Caucasian patients with end-stage renal disease; 3.35 standardized incidence ratio (95 percent CI: 2.59–4.40) for all female patients with end-stage renal disease in the study population (Stehman-Breen et al., 2000). For female patients ages 45 to 54 on dialysis in the study population, the observed/expected ratio was 20.0 (95 percent CI: 13.5–30.8) for hip fracture. For female patients ages 55 to 64 on dialysis in the study population, the observed/expected ratio was 10.2 (95 percent CI: 8.2–12.8) for hip fracture (Alem et al., 2000).
Gastric bypass	In patients with diabetes, gastric bypass had a hazard ratio of 1.26 (95 percent CI: 1.05–1.53) for risk of any type of fracture. In patients without diabetes, the hazard ratio was 1.32 (95 percent CI: 1.28–1.47) (Axelsson et al., 2018).
Hyperparathyroidism	Patients with primary hyperparathyroidism had a standardized incidence ratio of: • 3.2 (95 percent CI: 2.5–4.0) for vertebral fracture. • 2.2 (95 percent CI: 1.6–2.9) for distal forearm fracture. • 2.7 (95 percent CI: 2.1–3.5) for rib fracture. • 2.1 (95 percent CI: 1.2–3.5) for pelvic fracture (Khosla et al., 1999).
Lupus	Compared with similar-age women from a U.S. population sample, women ages 45 to 64 with lupus had a 7.6 standardized morbidity ratio for any self-reported fracture (95 percent CI: 5.1–10.7) (Ramsey-Goldman et al., 1999).
Marfan syndrome	No studies were identified assessing fracture risk in patients with Marfan syndrome. However, a large case-control study showed that patients with Marfan syndrome had lower bone mineral density compared with controls, independent of body mass index (Moura et al., 2006).

Exclusion	Risk of fracture
Osteogenesis imperfecta	Compared with a reference population, women with osteogenesis imperfecta had an incidence rate ratio of: • 5.9 (95 percent CI: 4.7–7.4) for any type of fracture in women 20 to 54 years old. • 8.0 (95 percent CI: 5.6–11.4) for any type of fracture in women 55 years old and older. • 1.6 (95 percent CI: 0.5–2.6) for spine fracture in women 55 years old and older. • 4.52 (95 percent CI: 2.79–6.26) for hip fracture in women 55 years old and older (Folkestad et al., 2017).
Psoriatic arthritis	Compared with controls, patients with psoriatic arthritis had a hazard ratio of 1.16 (95 percent CI: 1.06–1.27) for any type of fracture. For hip fracture, the hazard ratio was 1.17 (95 percent CI: 0.86–1.59). For vertebral fracture, the hazard ratio was 1.07 (95 percent CI: 0.66–1.72) (Ogdie et al., 2017).
Type 1 diabetes	Compared with controls, patients with type 1 diabetes are more likely to have a hip-fracture hospitalization (incidence rate ratio of 6.39; 95 percent CI: 1.94–22.35) and hip fracture (cause-specific hazard ratio of 7.11; 95 percent CI: 2.45–20.64) (Hamilton et al., 2017).

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

Initially, we constructed the measure to line up with the 2011 USPSTF recommendation and its supporting evidence. In April and May 2018, we developed a search string to capture literature focused on the overuse of DXA scans and searched PubMed for articles published since the release of the 2011 USPSTF guideline (January 2011 to January 2018). We searched for literature that addressed overuse of DXA scans in women under age 65 and also completed a clinical guideline scan for guidelines about DXA scans published in the United States, United Kingdom, and Canada.

To identify evidence for exclusions, we conducted a literature search for supplementary work to accompany the guidelines. The goal of the search was to identify independent factors that put a person at higher risk for fractures.

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

Alem AM, Sherrard DJ, Gillen DL, Weiss NS, Beresford SA, Heckbert SR, et al. Increased risk of hip fracture among patients with end-stage renal disease. Kidney Int. 2000 Jul;58(1):396-9.

Amarnath ALD, Franks P, Robbins JA, Xing G, Fenton JJ. Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study. J Gen Intern Med. 2015;30(12):1733-40. https://doi.org/10.1007/s11606-015-3349-8

American Academy of Family Physicians. Choosing Wisely: DEXA for osteoporosis recommendation. https://www.aafp.org/patient-care/clinical-recommendations/all/cw-osteoporosis.html. Accessed October 2, 2018.

Axelsson KF, Werling M, Eliasson B, Szabo E, Näslund I, Wedel H, et al. Fracture risk after gastric bypass surgery: a retrospective cohort study. J Bone Miner Res. 2018 Jul (published online ahead of print). doi: 10.1002/jbmr.3553

Dolan AL, Arden NK, Grahame R, Spector TD. Assessment of bone in Ehlers Danlos syndrome by ultrasound and densitometry. Ann Rheum Dis. 1998 Oct;57(10):630-3.

Fenton JJ, Robbins JA, Amarnath ALD, Franks P. Osteoporosis overtreatment in a regional health care system. JAMA Intern Med. 2016;176(3):391-3. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2478896

Folkestad L, Hald JD, Ersbøll AK, Gram J, Hermann AP, Langdahl B, et al. Fracture rates and fracture sites in patients with osteogenesis imperfecta: a nationwide register-based cohort study. J Bone Miner Res. 2017 Jan;32(1):125-34.

Hamilton EJ, Davis WA, Bruce DG, Davis TME. Risk and associates of incident hip fracture in type 1 diabetes: the Fremantle Diabetes Study. Diabetes Res Clin Pract. 2017 Dec;134:153-60.

Kanis JA, Johansson H, Oden A, Johnell O, de Laet C, Melton III L, et al. A meta-analysis of prior corticosteroid use and fracture risk. J Bone Miner Res. 2004 Jun;19(6):893-9.

Khosla S, Melton LJ, Wermers RA, Crowson CS, O'Fallon W, Riggs Bl. Primary hyperparathyroidism and the risk of fracture: a population-based study. J Bone Miner Res. 1999 Oct;14(10):1700-7.

Lasser EC, Pfoh ER, Chang HY, Chan KS, Bailey JC, Kharrazi H, et al. Has Choosing Wisely® affected rates of dual-energy X-ray absorptiometry use? Osteoporos Int. 2016;27(7):2311-6. https://doi.org/10.1007/s00198-016-3511-0

Messina C, Bandirali M, Sconfienza LM, D'Alonzo NK, Di Leo G, Papini GDE, et al. Prevalence and type of errors in dual-energy X-ray absorptiometry. Eur Radiol. 2015;25(5):1504-11. https://doi.org/10.1007/s00330-014-3509-y

Moura B, Tubach F, Sulpice M, Boileau C, Jondeau G, Muti C, et al. Bone mineral density in Marfan syndrome: a large case-control study. Joint Bone Spine. 2006 Dec;73(6):733-5.

Moynihan R, Sims R, Hersch J, Thomas R, Glasziou P, McCaffery K. Communicating about overdiagnosis: Learning from community focus groups on osteoporosis. PLoS ONE. 2017;12(2),1-16. https://doi.org/10.1371/journal.pone.0170142

Ogdie A, Harter L, Shin D, Baker J, Takeshita J, Choi HK, et al. The risk of fracture among patients with psoriatic arthritis and psoriasis: a population-based study. Ann Rheum Dis. 2017 May;76(5):882-5.

Ramsey-Goldman R, Dunn JE, Huang CF, Dunlop D, Rairie JE, Fitzgerald S, et al. Frequency of fractures in women with systemic lupus erythematosus: comparison with United States population data. Arthritis Rheum. 1999 May; 42(5):882-90.

Schnatz PF, Marakovits KA, Dubois M, O'Sullivan DM. Osteoporosis screening and treatment guidelines: are they being followed? Menopause. 2011;18:1072-8.

Silverman SL, Calderon AD. The utility and limitations of FRAX: a US perspective. Curr Osteoporos Rep. 2010;8(4):192-7. https://doi.org/10.1007/s11914-010-0032-1

Stehman-Breen CO, Sherrard DJ, Alem AM, Gillen DL, Heckbert SR, Wong CS, et al. Risk factors for hip fracture among patients with end-stage renal disease. Kidney Int. 2000 Nov;58(5):2200-5.

Tseng OL, Spinelli JJ, Gotay CC, Ho WY, McBride ML, Dawes MG. Aromatase inhibitors are associated with a higher fracture risk than tamoxifen: a systematic review and meta-analysis. Ther Adv Musculoskelet Dis. 2018 Apr;10(4):71-90.

Vestergaard P, Lindholm J, Jørgensen JO, Hagen C, Hoeck HC, Laurberg P, et al. Increased risk of osteoporotic fractures in patients with Cushing's syndrome. Eur J Endocrinol. 2002 Jan;146(1):51-6.

Weiss RJ, Wick MC, Ackermann PW, Montgomery SM. Increased fracture risk in patients with rheumatic disorders and other inflammatory diseases—a case-control study with 53,108 patients with fracture. J Rheumatol. 2010 Nov;37(11):2247-50.

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.

This measure is expected to increase the recording of patient risk for fracture data and to decrease the number of inappropriate DXA scans. Current osteoporosis guidelines recommend using bone measurement testing to assess osteoporosis risk in women ages 65 and older. In postmenopausal women younger than 65, guidelines recommend using a formal clinical risk assessment tool to establish patients' risk for osteoporosis in order to determine whether to screen them for osteoporosis using bone measurement testing. Clinical information such as age, BMI, parental history of hip fracture, smoking, and alcohol use can be used to determine a woman's fracture risk (U.S. Preventive Services Task Force, 2018).

In addition, there are potentially avoidable harms associated with screening for osteoporosis in general, including exposure to radiation, false-positive exams, and the side effects of unnecessary osteoporosis medications, which add costs to an already burdened health care system (Lim et al., 2009).

Citations:

Lim LS, Hoeksema LJ, Sherin K. Screening for osteoporosis in the adult U.S. population: ACPM position statement on preventive practice. Am J Prev Med. 2009;36(4):366-75.

U.S. Preventive Services Task Force. Screening for osteoporosis to prevent fractures: U.S. Preventive Services Task Force recommendation statement." JAMA. 2018;319(24):2521-31.

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

This measure has not yet been implemented and does not have performance data. However, from testing, we have an indication of performance scores based on 2013 encounters across 269 primary care providers (PCPs) at two sites: a primary care practice in suburban Michigan and a large multispecialty group in New York. (We also contracted with a third site, a large multispecialty group in Maryland. However, this site independently conducted analyses based on 2012 encounters and sent its results to measure developers. The site did not

provide clinician-level performance scores.) In addition, we have data from 2,508,693 female patients ages 50 to 64 who were covered by one large multistate health plan and had a DXA scan in 2012.

In data on 7.5 million women from one large health plan, 6.7 percent of the women ages 50 to 64 had potentially inappropriate DXA scans. Although these data could not be analyzed at the clinician level, we present them because they indicate how the measure might perform if implemented nationally. Please note that the claims analysis is based on DXA scans performed rather than on DXA scans ordered (as specified in the measure), so the numbers might be lower than they would be if the measure were implemented.

The clinician-level data presented below are from only two sites, and thus they may not be representative of national performance.

In EHR data from 269 PCPs at two sites, the rates of potentially inappropriate DXA scans varied from 0.0 to 100 percent. Performance was skewed left, with the top decile of performers (that is, the worst performers) ordering inappropriate DXA scans for at least 10 percent of patients in the denominator. These results suggest that about 10 percent of clinicians have room for improvement.

Among the 269 PCPs at the two sites, the performance rate statistics were as follows:

Mean: 3 percent

Standard deviation: 9 percent

Minimum: 0 percent

Maximum: 100 percent

Interquartile range: 0 to 0.5 percent

10th percentile: 0 percent

50th percentile: 0 percent

90th percentile: 10 percent

95th percentile: 19 percent

99th percentile: 33 percent

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.

Studies suggest that among women who have had a DXA scan, about 40 percent do not meet the risk factors for frailty (Schnatz et al., 2011). Studies also indicate that DXA scans are overused, albeit at low rates. A retrospective longitudinal analysis conducted across 34 practices showed no difference in the rates of DXA scan usage before and after the publication of the Choosing Wisely recommendation about DXA overuse; rates were 2.6 percent before and 2.0 percent after (Lasser et al., 2016).

In addition, a retrospective cohort study of 13 practices assessed the three-, five-, and seven-year incidence of inappropriate and appropriate DXA scans. This study revealed a three-year incidence of DXA scans of 18.4 percent in women ages 50 to 59 without osteoporosis risk factors, and 24.9 percent in women ages 60 to 64 without risk factors (Amaranth et al., 2015).

Citations:

Amarnath ALD, Franks P, Robbins JA, Xing G, Fenton JJ. Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study. J Gen Intern Med. 2015; 30(12):1733-40. https://doi.org/10.1007/s11606-015-3349-8

Lasser EC, Pfoh ER, Chang HY, Chan KS, Bailey JC, Kharrazi H, et al. Has Choosing Wisely® affected rates of dual-energy X-ray absorptiometry use? Osteoporos Int. 2016; 27(7):2311-6. https://doi.org/10.1007/s00198-016-3511-0

Schnatz PF, Marakovits KA, Dubois M, O'Sullivan DM. Osteoporosis screening and treatment guidelines: are they being followed? Menopause. 2011; 18:1072-8.

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.*

This measure has not yet been implemented and does not have performance data. To understand how performance on this measure varies by patient characteristics, we compared patient-level measure results by age and race in the three test sites for which we had EHR data. Two sites provided data from 2013 encounters; the third conducted its own analyses based on 2012 encounters and sent the results to the measure developers.

The results below summarize the rates of potentially inappropriate DXA scans by age and race from these sites. The rate was highest among women ages 60 and older across two sites (the third site merged results for women ages 50 to 64). At two sites, black women had significantly lower rates of potentially inappropriate DXA scans than white women. Please note that the results stratified by race were calculated using an earlier version of the measure that included women ages 18 to 64.

RATES ON POTENTIAL DXA-OVERUSE MEASURE, BY AGE AND SITE

Note: Rates were calculated using EHR extracts from three sites.

Site 1

Ages 50–59: 0.25 percent

Ages 60–64: 0.29 percent

Site 2 (Site 2 combined the data for patients ages 50 to 64 in a single age bracket.)

Ages 50–64: 5.70 percent

Site 3

Ages 50–59: 6.20 percent

Ages 60–64: 8.19 percent

Rates on potential DXA-overuse measure, by race and site

Note: Rates were calculated using EHR extracts from three sites for women ages 18 to 64.

Site 1

White—0.11 percent

Black—0.07 percent

Asian—0.12 percent

Other—0.05 percent

Missing—0.08 percent

Site 2

White—2.36 percent

Black—1.23 percent

Asian—2.43 percent

Other—4.87 percent

Missing—1.83 percent

Site 3

White—2.79 percent

Black—2.67 percent

Asian—1.76 percent

Other—1.72 percent

Missing—2.28 percent

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

The literature also suggests disparities between black and white women with regard to DXA scans. In a gender matched study on women ages 60 and older in primary care practices, only 29.8 percent of black women were referred for a DXA scan, compared with 38.4 percent of white women ($p < 0.05$) (Hamrick et al., 2012). Of the referred women, 20.8 percent of the black women had the scan, compared with 27.0 percent of the white women ($p < 0.05$) (Hamrick et al., 2012). Also, among included women with a diagnosis of osteoporosis, black women were less likely to receive medication (79.6 percent) than were white women (89.2 percent) ($p < 0.05$), controlling for both age and BMI. But there was no difference in the pattern of follow-up visits between the two races (Hamrick et al., 2012).

Although the literature shows that all ethnicities are at risk for osteoporosis, the prevalence of osteoporosis differs across races and ethnicities. In 2010, an estimated 15.8 percent of non-Hispanic white women, 7.7 of non-Hispanic black women, and 20.4 percent of Mexican American women had osteoporosis of femoral neck or lumbar spine (Wright et al., 2014). Understanding these differences among women of different ethnicities is helpful as we continue to look at DXA scans in the population.

Citations:

Hamrick I, Cao Q, Aqbafe-Mosley D, Cummings DM. Osteoporosis health care disparities in postmenopausal women. J Womens Health. 2012 Dec;21(12):1232-6.

Wright NC, Looker AC, Saag KG, Curtis JR, Delzell ES, Randall S, et al. The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine. J Bone Miner Res. 2014 Nov;29(11):2520-6.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

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

No link to the current specifications exist; the specifications are attached in accordance with Question S. 2a.

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

This is an eMeasure **Attachment:** AppropriateDXAScan_v5_5_Artifacts-636687330076328450.zip, CMS249v1_Bonnie_test_cases-636687330189610329.xlsx, cms249bonnie_-002-.docx

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:** CMS249_ValueSets.xlsx

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

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

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

Not an instrument-based measure

S.3.1. <u>For maintenance of endorsement:</u> 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. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable. This is a new measure.

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.*

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

Female patients who received an order for at least one DXA scan in the measurement period.

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)*

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

Female patients who received an order for at least one DXA scan in the measurement period

Please refer to the attached Measure Authoring Tool (MAT) output and value sets.

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

Female patients ages 50 to 64 years with an encounter during the measurement period.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

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

Female patients ages 50 to 64 years with an encounter during the measurement period

Please refer to the attached MAT output and value sets.

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

The measure excludes patients who have a combination of risk factors (as determined by age) or one of the independent risk factors.

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.)*

Exclude patients with a combination of risk factors (as determined by age) or one of the independent risk factors

Ages: 50-54 (>=4 combination risk factors) or 1 independent risk factor

Ages: 55-59 (>=3 combination risk factors) or 1 independent risk factor

Ages: 60-64 (>=2 combination risk factors) or 1 independent risk factor

COMBINATION RISK FACTORS [The following risk factors are all combination risk factors; they are grouped by when they occur in relation to the measurement period]:

The following risk factors may occur any time in the patient's history but must be active during the measurement period:

White (race)

BMI <= 20 kg/m2 (must be the first BMI of the measurement period)

Smoker (current during the measurement period)

Alcohol consumption (> two units per day (one unit is 12 oz. of beer, 4 oz. of wine, or 1 oz. of liquor))

The following risk factor may occur any time in the patient's history and must not start during the measurement period:

Osteopenia

The following risk factors may occur at any time in the patient's history or during the measurement period:

Rheumatoid arthritis

Hyperthyroidism

Malabsorption Syndromes: celiac disease, inflammatory bowel disease, ulcerative colitis, Crohn´s disease, cystic fibrosis, malabsorption

Chronic liver disease

Chronic malnutrition

Documentation of history of hip fracture in parent

Osteoporotic fracture

Glucocorticoids (>= 5 mg/per day) [cumulative medication duration >= 90 days]

INDEPENDENT RISK FACTORS (The following risk factors are all independent risk factors; they are grouped by when they occur in relation to the measurement period):

The following risk factors may occur at any time in the patient's history and must not start during the measurement period:

Osteoporosis

The following risk factors may occur at any time in the patient's history:

Gastric bypass

FRAX[R] ten-year probability of all major osteoporosis related fracture >= 8.4 percent

Aromatase inhibitors

Type I Diabetes

End stage renal disease

Osteogenesis imperfecta

Ankylosing spondylitis

Psoriatic arthritis

Ehlers-Danlos syndrome

Cushing's syndrome

Hyperparathyroidism

Marfan syndrome

Lupus

Please refer to the attached MAT output and value sets.

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

Not applicable. This measure does not use stratification.

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.)*

Refer to items S.4 to S.9 for details, S2.a for the eCQM specification, and S2.b for value sets.

1. Determine the denominator. Identify female patients ages 50 to 64 who had an encounter during the measurement period.
2. Remove exclusions. Identify patients who meet the exclusion criteria and remove them from the denominator (female patients who have a combination of risk factors, as determined by age, or one of the independent risk factors).
3. Determine the numerator. Identify patients in the denominator (after removing patients who meet the exclusion criteria) who received at least one DXA scan order during the measurement period.

4. Calculate measure performance. Compute performance as a proportion: numerator cases divided by (denominator minus exclusions).

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. It is based on a clinician's entire patient population.

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

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

Not applicable. This measure 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.

Electronic Health Records

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

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

Not applicable. This measure is not instrument-based. Data are collected from structured fields of eligible clinicians' electronic health records (EHRs).

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)*

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

Clinician : Individual

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. This measure is not a composite measure.

2. Validity – See attached Measure Testing Submission Form

CMS249_Testing_Attachment.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)

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

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

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

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

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

n.a.

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

Three sites provided electronic health records (EHR) data for women between the ages of 18 and 64 who had encounters with eligible clinicians (ECs) during the measurement period (measurement period was 2012 for Site 1; 2013 for Sites 2 and 3). We also used claims data from one large multistate health plan. We used claims data for female patients with dual-energy X-ray absorptiometry (DXA) orders in 2013. We used the claims data as an initial way to estimate the percentage of women receiving potentially inappropriate DXA scans before contracting with sites to do in-depth validity and reliability testing, as well as to initially estimate the prevalence of exclusions.

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

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

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

Starting in August 2013, we recruited and selected three testing sites (see Table 1 for site details). Using specifications defined by the measure developer, Site 1 conducted analyses and provided output to measure developers. Sites 2 and 3 provided raw data and developers completed analyses. Consequently, some results from Site 1 are presented differently—for example, results were provided for women ages 50–64 but not further stratified results for women ages 50–59 and 60–64. Furthermore, Site 1 only provided data to support a subset of EHR data analyses (see 1.7).

We pursued testing sites that captured data elements for the measure in their existing EHR workflows. As a result, we recruited sites that could be considered advanced EHR users, suggesting that they are unlikely to be representative of the broader field of clinicians who treat the population of interest. Our approach thus offers evidence that the measure concept is achievable but does not provide conclusive evidence regarding the ability of all EHR users to implement these measures.

Table 1. Testing site characteristics

Characteristics	**Testing site**		
	Site 1	**Site 2**	**Site 3**
State	**Maryland**	**New York**	**Michigan**
Encounter dates in EHR data	2012	2013	2013
EHR system	Centricity	Epic	NextGen
Overall EHR experience	7 years	12 years	6 years
Practice type and specialty mix	Large multispecialty group	Large multispecialty group	Family practice and internal medicine
Number of sites	35	75	12
Participation in quality programs	PQRS	PQRS, PCMH, local initiatives including those related to Choosing Wisely and appropriate ordering of radiology procedures	PQRS, eRX, PCMH

PCMH = patient centered medical home; PQRS = physician quality reporting system; eRX = Electronic Prescribing Incentive Program; EHR = electronic health record

We also used claims data from one large multistate health plan to calculate the frequency of denominator exclusions and the percentage of potentially inappropriate DXA scans among women ages 50–64. The data

included 7.5 million covered lives, 7.1 million of which were insured in commercial plans. The majority of the remaining lives were insured by Medicaid; the data included about 50,000 Medicare beneficiaries.

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

As described in 1.5, we recruited and subcontracted with three sites to collect patient-level EHR data for the measure. During initial testing, the measure included women ages 18–64. Based on results of testing and feedback from experts, we later restricted the measure to women ages 50–64. In 2014, sites provided data for women between the ages of 18 and 64 who had encounters with ECs during the calendar year measurement period (see Table 1). When possible, we report results for women ages 50–64, since that age range aligns with the current measure specification. Data from 87,242 patients were collected from Site 1, 102,593 patients from Site 2, and 25,899 from Site 3. Tables 2 and 3 summarize patients' distribution by age and race, respectively.

Table 2. Patients' age distribution

Age	All sites		Site 1		Site 2		Site 3	
	N	%	N	%	N	%	N	%
18–29	44,642	20.7	20,952	24.0	18,524	18.1	5,166	19.9
30–39	55,187	25.5	20,099	23.0	30,506	29.7	4,582	17.7
40–49	49,127	22.8	19,532	22.4	23,667	23.1	5,928	22.9
50–59	54,403	25.2	26,659[a]	30.6	20,473	20.0	7,271	28.1
60–64	12,375	5.7	–	–	9,423	9.2	2,952	11.4
Total	215,734	100.0	87,242	100.0	102,593	100.0	25,899	100.0

Source: Testing site EHR extracts sent to Mathematica.

Note: Due to rounding, some percentages on the total row do not sum to exactly 100 percent.

[a] Includes women ages 50–64.

Table 3. Patients' race distribution

Race	All sites		Site 1		Site 2		Site 3	
	N	%	N	%	N	%	N	%
White	114,329	53.0	49,346	56.6	45,583	44.4	19,400	74.9
Black	38,798	18.0	26,880	30.8	9,597	9.4	2,321	9.0
Asian	13,380	6.2	3,985	4.6	8,243	8	1,152	4.4
Other	20,781	9.6	4,763	5.5	15,870	15.5	148	0.6
Missing	28,446	13.2	2,268	2.6	23,300	22.7	2,878	11.1
Total	208,792	100.0	87,242	100.0	102,593	100.0	25,899	100.0

Source: Testing site EHR extracts sent to Mathematica.

Note: Due to rounding, some percentages on the total row do not sum to exactly 100 percent.

We also used claims data for 7.5 million women between the ages of 18 and 64 from one large health plan. Of these, 2,508,693 were between the ages of 50 and 64.

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.

We used EHR data from three sites to test the validity of the data elements and the frequency of denominator exclusions. We also tested ECs' performance score distribution and the measure's reliability at Sites 2 and 3.

We used claims data from one large health plan to test the frequency of denominator exclusions, and the frequency of inappropriate DXA scans among women ages 50–64.

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

We did not test social risk factors because none were available in the EHR or claims 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*)

To estimate the measure's reliability, we used a random split-half correlation approach supplemented with bootstrapping. This approach included the following steps. First, we split patients randomly into two groups. For each group, we calculated the average performance rate per EC. With the EC as the unit of analysis, we then estimated the Pearson correlation coefficient to measure the strength of association between the two rates, using a resampling (bootstrapping) technique to increase the precision of the estimate. The resampling repeated these steps 2,500 times, with the average correlation calculated across iterations. We considered reliability coefficients of 0.70 and higher satisfactory (Nunnally and Bernstein 1994). For each tested measure, we also assessed the proportion of variance in EC performance scores attributable to EC performance, which we calculated by squaring the reliability estimate.

We limited our sample in the reliability analysis to primary care physicians (PCPs) who ordered DXA scans during the measurement period for female patients ages 50–64. Sites included a provider type code in the EHR data reports, which we used to identify PCPs.

We tested reliability across different denominator thresholds because prior work had shown reliability is dependent on the number of denominator cases (Scholle et al. 2008).

References

Nunnally, J.C., and I.H. Bernstein. *Psychometric Theory*. New York: McGraw-Hill, 1994.

Scholle, S.H., J. Roski, J.L. Adams, D.L. Dunn, E.A. Kerr, D.P. Dugan, and R.E. Jensen. "Benchmarking Physician Performance: Reliability of Individual and Composite Measures," *Am J Manag Care*, vol. 14, no. 12, Dec. 2008, pp. 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)

Table 4 displays the results of the reliability analysis. The average reliability coefficient among PCPs with at least 20 patients ages 50–64 was 0.82.

Table 4. Reliability results

Denominator threshold	Number of PCPs	Number of patients	Average	Min	25th Percentile	Median	75th Percentile	Max
1 or more patients	269	19,162	0.25	0.02	0.13	0.24	0.34	0.60
10 or more patient	170	18,791	0.68	0.35	0.61	0.69	0.76	0.91
20 or more patients	138	18,370	0.82	0.64	0.80	0.83	0.85	0.91

Source: Rates were calculated using EHR data from Sites 2 and 3. Site 1 did not conduct a reliability analysis.

Note: Reliability analysis restricted to primary care physicians and patients between the ages of 50 and 64.

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

The results indicate that for ECs with 20 or more patients in the denominator, the measure is reliable, with a median reliability of 0.83 and an interquartile range of 0.80–0.85. Measures with reliability coefficients of 0.70 are generally considered adequately reliable (Nunnally and Bernstein 1994). The lowest reliability estimate among the total group of 138 PCPs with at least 20 patients ages 50–64 in the denominator was 0.64 and the reliability estimate for the first percentile among this group was 0.72. These results suggest that the vast majority of PCPs were close to or above the reliability threshold of 0.70. The measure was reliable for about half of the PCPs in our sample, with 10 or more patients eligible for the denominator.

Reference

Nunnally, J. C., and I.H. Bernstein. *Psychometric Theory*. New York: McGraw-Hill 1994.

_____

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 <u>performance measure score</u> 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).*

To ensure data used in calculating a measure score accurately reflected the care a patient received, such as whether she received a DXA scan or had risk factors for osteoporotic fracture, we assessed validity at the data element level. We drew a random sample of patients and extracted data elements from their EHR records (EHR extract), which we compared with data manually abstracted through a detailed, visual review of the patients' EHR (manual abstract). Using an a priori power analysis, we determined that each site needed to abstract a minimum of 200 charts per measure to achieve 80 percent power to detect statistically significant differences between manually abstracted and EHR extracted data. We manually abstracted data for 200 patients for each measure at Sites 2 and 3. Clinical staff at Site 1 were responsible for abstracting data for 216 patients. We assessed validity using kappa agreement statistics to estimate the chance-adjusted agreement between the two data sources for the sampled patients at each site. This approach allowed us to assess the

validity of the EHR extract against a definitive record of the patients' care and to report overall agreement, sensitivity, and specificity. We then stratified validity results by site to obtain an understanding of how site characteristics (for example, documentation patterns) affected data element validity.

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

Chance-adjusted agreement between sites' EHR extracts and manually abstracted data for the numerator condition (DXA order) was 0.91, -0.01, and 0.93 at the three respective sites (Table 5). We also calculated chance-adjusted agreement for denominator exclusions; however, given the low prevalence of these data elements, these results are not reliable.

Table 5. Agreement between chart abstracted data and EHR extract

Site	Measure element	Data element	Chart prevalence	EHR prevalence	Kappa	Overall agreement	Sensitivity	Specificity
1	Numerator	DXA order	30.6%	33.3%	0.91	96.3%	98.5%	95.3%
1	Exclusion	Chronic malnutrition	0.0%	0.5%	N/A	99.5%	N/A	99.5%
1	Exclusion	Marfan syndrome	0.5%	0.0%	N/A	99.5%	0.0%	100.0%
1	Exclusion	Ehlers-Danlos syndrome	0.0%	0.0%	N/A	100.0%	N/A	100.0%
1	Exclusion	Osteopenia	13.9%	7.9%	0.60	92.1%	50.0%	98.9%
1	Exclusion	Osteoporosis	5.6%	0.0%	N/A	94.4%	0.0%	100.0%
1	Exclusion	Prior osteoporotic fracture	0.0%	0.0%	N/A	100.0%	.	100.0%
1	Exclusion	Ankylosing spondylitis	0.5%	0.0%	N/A	99.5%	0.0%	100.0%
1	Exclusion	Lupus	1.4%	0.5%	0.50	99.1%	33.3%	100.0%
1	Exclusion	Rheumatoid arthritis	0.5%	0.5%	1	100.0%	100.0%	100.0%
1	Exclusion	Type 1 diabetes	0.5%	0.5%	1	100.0%	100.0%	100.0%
1	Exclusion	Hyperthyroidism	0.5%	0.5%	1	100.0%	100.0%	100.0%
1	Exclusion	Hyperparathyroidism	–	–	–	–	–	–
1	Exclusion	Cushing's syndrome	0.0%	0.0%	N/A	100.0%	N/A	100.0%
1	Exclusion	Malabsorption syndrome	0.9%	0.5%	-0.01	98.6%	0.0%	99.5%
1	Exclusion	Chronic liver disease	1.4%	1.4%	0.32	98.1%	33.3%	99.1%
1	Exclusion	End-stage renal disease	0.0%	0.0%		100.0%	.	100.0%
1	Exclusion	Psoriatic arthritis	0.5%	0.5%	1	100.0%	100.0%	100.0%
1	Exclusion	Gastric bypass	–	–	–	–	–	–
1	Exclusion	Glucocorticoids	–	–	–	–	–	–
1	Exclusion	Risk of osteoporotic fracture[1]	–	–	–	–	–	–
1	Exclusion	Smoker	86.1%	83.3%	0.82	95.4%	95.7%	93.3%
2	Numerator	DXA order	48.5%	0.5%	-0.01	51.0%	0.0%	99.0%
2	Exclusion	Chronic malnutrition	0.0%	0.5%	N/A	99.5%	N/A	99.5%
2	Exclusion	Marfan syndrome	0.0%	0.0%	N/A	100.0%	N/A	100.0%
2	Exclusion	Ehlers-Danlos syndrome	0.0%	0.0%	N/A	100.0%	N/A	100.0%
2	Exclusion	Osteopenia	17.5%	1.5%	-0.03	81.0%	0.0%	98.2%
2	Exclusion	Osteoporosis	10.0%	1.5%	-0.03	88.5%	0.0%	98.3%
2	Exclusion	Prior osteoporotic fracture	3.5%	0.0%	N/A	96.5%	0.0%	100.0%
2	Exclusion	Ankylosing spondylitis	0.0%	0.0%	N/A	100.0%	.	100.0%
2	Exclusion	Lupus	1.0%	0.0%	N/A	99.0%	0.0%	100.0%
2	Exclusion	Rheumatoid arthritis	2.0%	0.0%	N/A	98.0%	0.0%	100.0%
2	Exclusion	Type 1 diabetes	0.0%	0.5%	N/A	99.5%	.	99.5%
2	Exclusion	Hyperthyroidism	2.0%	1.0%	-0.01	97.0%	0.0%	99.0%

Site	Measure element	Data element	Chart prevalence	EHR prevalence	Kappa	Overall agreement	Sensitivity	Specificity
2	Exclusion	Hyperparathyroidism	1.5%	0.0%	N/A	98.5%	0.0%	100.0%
2	Exclusion	Cushing's syndrome	0.0%	0.0%	N/A	100.0%	N/A	100.0%
2	Exclusion	Malabsorption syndrome	3.0%	2.0%	-0.02	95.0%	0.0%	97.9%
2	Exclusion	Chronic liver disease	5.0%	1.5%	-0.02	93.5%	0.0%	98.4%
2	Exclusion	End-stage renal disease	0.5%	1.5%	-0.01	98.0%	0.0%	98.5%
2	Exclusion	Psoriatic arthritis	0.5%	0.0%	N/A	99.5%	0.0%	100.0%
2	Exclusion	Gastric bypass	0.0%	0.0%	N/A	100.0%	N/A	100.0%
2	Exclusion	Glucocorticoids	2.5%	0.0%	N/A	97.5%	0.0%	100.0%
2	Exclusion	Risk of osteoporotic fracture[1]	0.5%	0.0%	N/A	99.5%	0.0%	100.0%
2	Exclusion	Smoker	9.0%	7.0%	0.12	87.0%	16.7%	94.0%
3	Numerator	DXA order	11.0%	12.5%	0.93	99%	100%	98%
3	Exclusion	Chronic malnutrition	0.5%	0.5%	1.00	100%	100%	100%
3	Exclusion	Marfan syndrome	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Ehlers-Danlos syndrome	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Osteopenia	12.5%	7.5%	0.56	92%	48%	98%
3	Exclusion	Osteoporosis	3.5%	0.0%	N/A	97%	0%	100%
3	Exclusion	Prior osteoporotic fracture	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Ankylosing spondylitis	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Lupus	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Rheumatoid arthritis	1.0%	1.5%	0.80	100%	100%	99%
3	Exclusion	Type 1 diabetes	0.5%	0.5%	-0.01	99%	0%	99%
3	Exclusion	Hyperthyroidism	1.0%	0.0%	N/A	99%	0%	100%
3	Exclusion	Hyperparathyroidism	1.5%	0.5%	0.50	99%	33%	100%
3	Exclusion	Cushing's syndrome	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Malabsorption syndrome	3.0%	1.5%	0.66	99%	50%	100%
3	Exclusion	Chronic liver disease	1.5%	2.0%	0.86	100%	100%	99%
3	Exclusion	End-stage renal disease	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Psoriatic arthritis	0.0%	0.0%	N/A	100%	.	100%
3	Exclusion	Gastric bypass	0.5%	0.5%	1.00	100%	100%	100%
3	Exclusion	Glucocorticoids	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Risk of osteoporotic fracture[1]	0.0%	0.0%	N/A	100%	N/A	100%
3	Exclusion	Smoker	9.0%	9.5%	0.97	100%	100%	99%

Source: Results comparing EHR extracted data and manually abstracted EHR chart data from three sites in 2013–2014. Results are based on 200 records from Sites 2 and 3 and 216 records from Site 1.

Dashes (–) in this table indicate elements that Site 1 was unable to calculate results based on their provision of incomplete information.

Note: The following data elements in the current specification are not included in the results: race, body mass index (BMI), alcohol consumption, osteogenesis imperfecta, aromatase inhibitors, documentation of hip fracture in parent.

[1] Measure specification used in testing did not include a data element focused on the FRAX® 10-year probability of osteoporotic fracture. Instead, it excluded women with a ten-year probability of osteoporotic fracture >=20 percent without specifying a risk assessment tool. After testing, we added a data element to the measure's specification which defined risk of osteoporotic fracture using a FRAX® 10-year probability of

osteoporotic fracture >=9.3 percent which aligned with the USPSTF recommendation on osteoporosis screening. The measure specification being submitted to NQF uses a FRAX® 10-year probability of >=8.4 percent, in alignment with the June 2018 USPSTF recommendation on osteoporosis screening.

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?*)

In general, we considered kappa scores below 0.40 to be indicative of poor agreement, scores of 0.40 to 0.75 to be indicative of intermediate to good agreement, and scores above 0.75 to be indicative of excellent agreement (Fleiss 1981).

Chance-adjusted agreement between sites' EHR extracts and manually abstracted data for the numerator condition (DXA order) was excellent at two of the three sites that participated in our testing (Sites 1 and 3). Staff at these sites mentioned during site visits that their physicians routinely used structured fields to capture orders for DXA scans. Site 2 had agreement equal to chance for DXA orders, which was attributable to a lack of EHR documentation for DXA scans in structured fields (0.5 percent in the EHR extract versus 48.5 percent in the manual abstract). We also calculated agreement between denominator exclusion data elements. However, due to low prevalence of these data elements, kappa results are not reliable. The most prevalent denominator exclusion (current smoker status) had very good kappa agreement at Sites 2 and 3 (0.82 and 0.97, respectively).

We did not test the data element validity of the 10-year probability of osteoporotic fracture because it was unavailable at all test sites. However, the data element is derived from the FRAX, a validated tool.

Reference

Fleiss, J.L. *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons, Inc., 1981.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b3

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

We explored the EHR records and claims data for the prevalence of exclusions among all women in the denominator. Women were excluded from the measure if they had at least one of the following conditions and/or met any of these criteria: osteopenia, osteoporosis, chronic liver disease, malabsorption syndrome, hyperthyroidism, rheumatoid arthritis, type I diabetes, lupus, chronic malnutrition, prior osteoporotic fracture, use of glucocorticoids, hyperparathyroidism, psoriatic arthritis, end-stage renal disease, ankylosing spondylitis, recent gastric bypass surgery, Cushing's syndrome, Ehlers-Danlos syndrome, Marfan syndrome, osteogenesis imperfecta, low BMI (≤20 kg/m^2), current smoker, high alcohol consumption (more than 2 units per drinking day, where one unit is 12 oz. of beer, 4 oz. of wine, or 1 oz. of liquor), or 10-year risk of osteoporotic fracture ≥9.3 percent.[1] We did not measure the prevalence of risk of osteoporotic fracture because sites did not include the variable in structured fields of their EHRs, nor were the data available in claims. Overall, prevalence rates for most exclusions were typically under 5 percent, with the exceptions of osteoporosis and osteopenia. Claims data show the prevalence of exclusions in all women ages 18 to 64, because that was the population included in the measure at the time of the analysis. However, we limited EHR extracts to women who were ages 50 to 64, to conform with the measure specification. Several exclusions were not available in claims data, such as BMI, smoking status, and alcohol consumption.

---

[1] The measure specification we submitted to NQF uses a FRAX® 10-year probability of >=8.4 percent, in alignment with the June 2018 USPSTF recommendation on osteoporosis screening.

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

Table 6 displays the prevalence of excluded data elements in claims and EHR data.

Table 6. Prevalence of measure exclusions

Exclusion	Claims (women ages 18–64 with DXA order)	Site 1 (women ages 50–64)	Site 2 (women ages 50–64)	Site 3 (women ages 50–64)
Osteopenia	46.9%	57.2%	4.3%	11.0%
Osteoporosis	26.2%	#	2.7%	0.5%
Chronic liver disease	4.9%	15.0%	1.9%	1.6%
Malabsorption syndrome	3.9%	8.1%	2.6%	1.7%
Rheumatoid arthritis	3.3%	6.9%	1.0%	1.1%
Hyperthyroidism	2.5%	3.3%	1.1%	#
Type I diabetes	1.5%	7.0%	0.6%	0.5%
Lupus	1.3%	3.4%	0.7%	#
Prior osteoporotic fracture	1.3%	–	#	#
Chronic malnutrition	0.5%	1.5%	#	#
Hyperparathyroidism	1.3%	–	0.7%	#
Glucocorticoids (oral only)	0.8%	–	–	–
Psoriatic arthritis	#	1.1%	#	#
End-stage renal disease	#	1.0%	1.1%	1.2%
Ankylosing spondylitis	#	#	#	#
Gastric bypass surgery	#	–	#	#
Cushing's Syndrome	#	#	#	#
Ehlers-Danlos syndrome	#	#	#	#
Marfan syndrome	#	#	#	#
Osteogenesis imperfecta	#	#	#	#
BMI <=20 kg/m^2	–	–	16.4%	9.8%
Current smoker	–	–	7.5%	12.0%
>2 units of alcohol per drinking day	–	–	5.5%	–

Source: Claims from one large health plan and testing site EHR extracts.

#: Prevalence was <0.5%

Dashes (–) in this table indicate data is not available. BMI, smoking status, and alcohol consumption are not generally available in claims data. Site 1 did not provide all data elements.

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*)

Using claims data, we examined the rate of denominator exclusions among women with DXA scans during measure testing. Many denominator exclusion conditions had extremely low prevalence (less than 2 percent). The three most prevalent conditions were osteopenia (47 percent), osteoporosis (26 percent), and chronic liver disease (4.9 percent). The prevalence of exclusions in EHR data varied across sites. Site 1 had a higher prevalence of risk factors than Sites 2 and 3, including osteopenia, chronic liver disease, malabsorption

syndrome, rheumatoid arthritis, and lupus. Sites 2 and 3 had a fairly high prevalence of patients with low BMI or current smokers, both risk factors that can exclude patients from the measure if they co-occur with other risk factors.

We did not test the exclusion of FRAX® risk of osteoporotic fracture >=9.3 percent[2] because this data element was added to the specification after our 2013–2014 testing. Discussions with testing sites during 2018 feasibility testing suggest that practices rarely capture the FRAX® score in structured fields of their EHR, so we would expect prevalence of this data element to be low as well. However, the measure specification excludes patients based on combination and independent risk factors that serve as inputs to the FRAX® tool. Therefore, although use of the FRAX® score can facilitate ECs' identification of patients to exclude from the measure, the specifications provide an alternative way to identify these patients.

Although many exclusions have low prevalence, they are based on evidence and add to the face validity of the measure. Therefore, we retained them in the measure. Some risk factors, such as osteogenesis imperfecta, are rare and likely to be infrequent in PCP data but could be much more prevalent for specialists who chose to report this measure. Furthermore, variation across sites' EHR data demonstrate that risk factors might be more prevalent in some settings than others. Therefore, the exclusions are important for ensuring that practices' performance scores are based on patients lacking risk factors for whom DXA scans might be unnecessary.

_____

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

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4***.***

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

☒ No risk adjustment or stratification

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

n.a. This measure does not use risk adjustment or stratification.

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.

n.a. This is a process measure.

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?

n.a.

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

☐ **Published literature**

[2] The measure specification we submitted to NQF uses a FRAX® 10-year probability of >=8.4 percent, in alignment with the June 2018 USPSTF recommendation on osteoporosis screening.

☐ **Internal data analysis**

☐ **Other (please describe)**

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

n.a.

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.**

n.a.

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

n.a.

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

n.a.

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

n.a.

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

n.a.

2b3.9. Results of Risk Stratification Analysis:

n.a.

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*)

n.a.

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*)

n.a.

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)*

Using EHR data from Sites 2 and 3, we computed and examined the distribution of performance scores for 269 PCPs (Table 7), which represented one type of EC that might choose to report this measure using EHR data. (Site 1 did not provide EC performance score distributions). Using claims data, we also calculated the percentage of potentially inappropriate DXA scans among 2,508,693 women ages 50–64 who were insured by one large health plan.

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*)

Among 2,508,693 women ages 50–64 insured by one large health plan, 6.7 percent had potentially inappropriate DXA scans as defined by the measure.

Table 7 displays the performance score distribution calculated from EHR data from two sites.

Table 7. Performance distribution

EP type	Number of EPs	50th percentile	75th percentile	90th percentile	95th percentile	99th percentile
PCPs (patients ages 50–64)	269	0.0%	0.5%	10.0%	19.2%	33.3%
PCPs (10 or more patients ages 50–64)	170	0.0%	1.6%	6.1%	14.3%	22.2%
PCPs (20 or more patients age 50–64)	140	0.0%	2.2%	6.4%	11.3%	21.0%

Note: Rates were calculated using EHR extracts from Sites 2 and 3. Lower scores indicate higher quality.

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 claims results demonstrate that there is an opportunity to reduce the number of women ages 50–64 receiving inappropriate DXA scans, with nearly 7 percent of women receiving a potentially inappropriate DXA scan.

The distribution from EHR data was skewed to the left (median performance was 0.0 percent), suggesting many PCPs were not ordering potentially inappropriate DXA scans for their patients. Among PCPs with 20 or more patients ages 50–64, PCPs in the highest decile of the distribution (that is, the poorest decile of physician performance) had performance scores between about 6 and 21 percent, similar to results for PCPs with 10 or more patients in the denominator.

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*)

n.a. This measure uses one set of specifications.

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*)

n.a.

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*)

n.a.

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*)

We did not explicitly test the distribution of missing data for this measure; however, our data file construction and data element validity results inform our understanding of the potential for systematic bias due to missing data. When comparing the data extracted from the EHR with a manual review of the full medical record, data element validity testing can be used to inform the level of missingness for individual data elements. Missing data will result in low overall agreement and chance-adjusted agreement (kappa).

The testing was limited to patients ages 50 to 64 eligible for the measure's denominator. In the data files submitted by test sites, there was no distinction between a negative (for example, confirmation that the patient was diagnosed with osteoporosis) and missing data. Where sites reported data for at least one patient, we assumed that blank records indicated no relevant data for those patients. For example, we assumed a patient with no data indicating osteoporosis did not have osteoporosis; we did not exclude that patient from the denominator based on lack of data regarding osteoporosis.

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)

As described above, data element validity testing, when comparing the data extracted from the EHR with a complete chart review for a sample of patients, can be used to inform the level of missingness for individual data elements. Missing data will result in low overall and chance-adjusted agreement (kappa). Lack of missing data will result in high overall and chance-adjusted agreement.

As shown in Table 5, agreement between the EHR extract and an abstract of the full chart showed high agreement and a lack of systematic missing information. Of the 22 data elements tested, all had overall agreement rates greater than 90 percent at two or more sites. Only four data elements scored less than 90 percent at Site 2, and three of these data elements scored above 80 percent. Due to low prevalence of many excluded data elements, we could not calculate kappa at all sites for all data elements. However, for DXA orders, the data element necessary to calculate the numerator, kappa was >.95 at two of the three sites. Smoking status, one of the most prevalent exclusions, had kappa >0.8 at two of the three sites. Osteoporosis and osteopenia, two additional exclusion data elements that were fairly prevalent, had kappa >0.55 at two of the three sites. Variation across testing sites indicates that missingness for these data elements was not systematic.

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*)

As noted above, we interpreted data missing from numerator fields as indicating that the patient did not receive the service, and data missing from denominator exclusion fields as indicating that the patient should be included in the measure.

Based on our analysis of data element validity showing site-level variation in kappa and overall agreement rates for the four data elements with lowest EHR/chart agreement (overall agreement rates < 90 percent and kappa <55 percent for prevalent data elements), we conclude that missing data for these data elements is not systematic.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

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

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

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

Attachment: DXA_Feasibility_Scorecard_-1-.xlsx,DXA_Feasibility_Narrative_Final.docx

3c. Data Collection Strategy

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

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

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

This is a new measure that has not yet been implemented. Attached to this submission are two documents—a feasibility summary and scorecard—that describe the difficulties regarding data collection.

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)*.

The FRAX may be accessed online for free. Clinicians can also purchase a desktop version if desired. To our knowledge, there are no fees, licensing, or other requirements associated with using any other aspect of the measure as specified, such as the value or code sets, programming code, or algorithm. The measure is available for public use.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting	Payment Program CMS Merit-based Incentive Payment System (MIPS) https://qpp.cms.gov/mips/quality-measures

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

In the final CY2019 Medicare Physician Fee Schedule rule posted on November 1, 2018, CMS added this measure to MIPS beginning with performance period 2019. MIPS streamlines three historical Medicare programs – the Physician Quality Reporting System, the Value-based Payment Modifier Program, and the Medicare Electronic Health Record Incentive Program – into a single payment program as part of CMS efforts to move clinicians to a performance-based payment system. MIPS is a national program where eligible clinicians can choose to report quality measures most meaningful to their practice. Clinicians will have the option to report this measure in 2019.

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?*)

There are no reasons, such as policies or accessibility, which prohibit the use of this measure. CMS has adopted this measure for use in its MIPS program for performance period 2019 and future years. More information can be found in Section 4a1.3.

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.)

CMS submitted the measure to the Measures Under Consideration list for MIPS in June 2017. The Measure Applications Partnership reviewed the measure in December 2017 and recommended the measure for inclusion in the program with conditional support (pending NQF endorsement). CMS adopted this measure for use in its MIPS program for performance period 2019 and future years.

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.

This is a new measure and has not yet been implemented, so we have not shared performance results with the entities being measured. However, as part of measure development and testing, we computed measure performance for two physician practices. We shared their performance data with various organizations or individuals, as described below.

With the two physician practices (test sites), we shared their individual performance rates but did not share the performance rates of the other test sites. The clinicians at these sites are the types of eligible clinicians who may, in the future, report on this measure as part of the MIPS program.

We shared performance data from both test sites with a technical expert panel (TEP). The TEP consisted of health system representatives, EHR vendors, patients, consumer representatives, and clinicians. It included clinicians who may, in the future, report on this measure as part of the MIPS program, along with other experts who would not report on this measure (for example, EHR vendors who do not work in a clinician practice).

We also shared performance data from both test sites with a DXA Overuse expert work group (EWG). The EWG consisted of experts in osteoporosis, skeletal health, and overuse measurement. It included clinicians who may, in the future, report on this measure as part of the MIPS program, along with other experts who would not report on this measure (for example, measure development experts who do not work in a clinician practice).

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.

This is a new measure that CMS has not yet implemented, so we do not have national performance results to share. However, during measure development and testing, we computed measure performance for two clinician practices and shared the results with the test sites, the TEP, and the EWG. With each test site, we shared only the overall measure performance for that practice. With the TEP and EWG, we shared de-identified overall measure performance across the two test sites. We shared these data once with each group. During the meetings in which we shared the data, we also reviewed the measure specifications.

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.

During measure testing, we gave the test sites an opportunity to discuss any questions or concerns they had about their measure performance.

During our meetings with the TEP and EWG, we gave the members an opportunity to discuss any questions or concerns they had about the shared performance information.

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

The two test sites did not share any significant concerns about their performance on the measure.

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

The TEP and EWG did not share any significant concerns about clinician performance on the measure.

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.

The feedback described above did not result in changes to the measure specifications.

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.

The intent of this measure is to decrease the use of DXA scans among people who are at low risk for osteoporotic fracture, thereby reducing DXA-related harms. Although the measure is not yet in use, we expect that its implementation will improve quality of care by helping clinicians track their performance and by motivating them to reduce the number of inappropriate DXA scans they order.

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 that CMS has not yet implemented in a program. When CMS implements the measure, it could cause women ages 50 to 64 with osteoporosis who do not have the risk factors identified in the measure—or who have the risk factors but not the number specified—to miss needed DXA screenings. Also, the applicability of the FRAX to nonwhite subgroups has not yet been widely studied (Viswanathan et al., 2018). Nonwhite women and women with risk factors other than those identified by the measure could fail to begin or experience unnecessary delays in appropriate treatment for osteoporosis.

Citation:

Viswanathan M, Reddy S, Berkman N, Cullen K, Middleton J, Nicholson W, et al. Screening to prevent osteoporotic fractures: updated evidence report and systematic review for the U.S. Preventive Services Task Force." JAMA. 2018;319(24):2532-51.

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

This measure does not explicitly assess clinician use of clinical risk assessment tools to determine patient risk for osteoporotic fracture (as recommended by the U.S. Preventive Services Task Force). However, it will encourage the use of those tools—particularly the FRAX—because clinicians will notice its inclusion in the measure as a method for identifying patients at high risk for fracture; clinicians may decide that this tool is an efficient way to screen patients before ordering a DXA scan. The measure could also increase clinicians' consistency in determining which patients are at high risk for osteoporotic fracture—and therefore eligible for a DXA scan.

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)

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

Yes

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

(NQF 0046) Screening or Therapy for Osteoporosis for Women Aged 65 Years and Older: Percentage of female patients aged 65-85 years of age who ever had a central dual-energy X-ray absorptiometry (DXA) to check for osteoporosis. NQF 0046 is in MIPS and is specified for claims and registry reporting. It complements the proposed measure because it assesses the percentage of women who receive an appropriate osteoporosis screening after age 65. There are some differences between the measures, but these are appropriate based on the measures' intents. NQF 0046 assesses for documentation of DXA results, whereas the proposed measure assesses for DXA orders. Assessing for DXA orders makes sense because the proposed measure focuses on overuse of DXA screening. Also, NQF 0046 is limited to DXA scans of the hip or spine (that is, central DXA scans), whereas the proposed measure assesses for central and peripheral DXA scans. In its 2011 recommendation, the U.S. Preventive Services Task Force recommended using central DXA scans to assess for osteoporosis—and NQF 0046 complies with this recommendation. But the proposed measure, as an overuse measure, assesses for any type of DXA scan because any type could be inappropriate. Together, these two measures assess the appropriate use of DXA scans in women 65 and older, along with inappropriate use of DXA scans in women under age 65.

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. We did not identify any competing measures.

Appendix

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

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services, Center for Clinical Standards and Quality, Quality Measurement and Value-Based Incentives Group (QMVIG), Division of Electronic and Clinician Quality, MS S3-02-01

Co.2 Point of Contact: Susan, Arday, B.S.P.H., M.H.S., C.H.E.S., Susan.Arday@cms.hhs.gov, 410-786-3141-

Co.3 Measure Developer if different from Measure Steward: NCQA

Co.4 Point of Contact: Jenna, Williams-Bader, bader@ncqa.org, 202-955-5103-

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.

The following individuals participated in the DXA Overuse EWG. We selected EWG members based on their expertise in osteoporosis, skeletal health, and overuse measurement. They provided feedback throughout the measure's development, from 2013 to 2014—commenting on the clinical components of the measure, including the denominator, numerator, and exclusions, and on the measure's importance, feasibility, validity, and usability.

Itara Barnes, Medical University of South Carolina

Meryl S. LeBoff, M.D., Brigham and Women's Hospital

Michael LeFevre, M.D., M.S.P.H., University of Missouri

Mark Robbins, M.D., Harvard Vanguard

Kenneth Saag, M.D., M.Sc., University of Alabama at Birmingham

The following individuals participated in the TEP. This multistakeholder group had representatives from health systems, clinician practices, EHR vendors, and consumer advocacy organizations. The TEP provided feedback throughout the measure's development, from 2013 to 2014, on the importance, feasibility, validity, and usability of the measure.

Ayodola Anise, M.H.S., senior research associate, Engelberg Center for Health Care Reform, The Brookings Institute

Jessica Bartell, M.D., M.S., clinical informatics physician, Epic

Nate Bennett, M.D., physician, Preferred Primary Care Physicians

Jason Colquitt, executive director, research services, Greenway Medical Technologies, Inc.

William F. Groneman, M.H.A., executive vice president, system development, TriHealth, Inc.

Erin A. Mackay, M.P.H., associate director, health information technology systems, National Partnership for Women & Families

Jon D. Morrow, M.D., M.B.A., M.A., F.A.C.O.G, executive vice president, system development, General Electric Healthcare

Daniel Todd Rosenthal, M.D., M.Sc., M.P.H., director of health care intelligence, Inova Health Systems

Shannon Sims, M.D., Ph.D., director of clinical informatics and medical director of information services, Rush University Medical Center

Samuel S. Spicer, M.D., M.M.M., vice president of medical affairs, New Hanover Regional Medical Center

Rachelle "Shelly" Spiro, R.Ph., F.A.S.C.P., director, Pharmacy e-Health Information Technology Collaborative

Andy Steele, M.D., M.P.H, M.Sc., director of medical informatics, Denver Health

Jonathan P. Weiner, Dr.P.H., M.S., professor and program director, Johns Hopkins Bloomberg School of Public Health

Thomas R. Williams, M.P.H., M.B.A., Dr.P.H., executive director, Integrated Healthcare Association

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2018

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure? CMS conducts an annual review to determine potential updates to the measure.

Ad.5 When is the next scheduled review/update for this measure? 2019

Ad.6 Copyright statement: This Physician Performance Measure (Measure) and related data specifications are owned and stewarded by the Centers for Medicare & Medicaid Services (CMS). This measure was developed under CMS Contract No. HHSM-500-2013-13011I, Task Order HHSM-500-T00001. Mathematica Policy Research and the National Committee for Quality Assurance (NCQA) supported development of this electronic measure. NCQA is not responsible for any use of the Measure. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures and NCQA has no liability to anyone who relies on such measures or specifications.

Limited proprietary coding is contained in the Measure specifications for user convenience. Users of proprietary code sets should obtain all necessary licenses from the owners of the code sets. NCQA disclaims all liability for use or accuracy of any third party codes contained in the specifications.

CPT(R) contained in the Measure specifications is copyright 2004-2018 American Medical Association. LOINC(R) copyright 2004-2018 Regenstrief Institute, Inc. This material contains SNOMED Clinical Terms(R) (SNOMED CT[R]) copyright 2004-2018 International Health Terminology Standards Development Organisation. ICD-10 copyright 2018 World Health Organization. All Rights Reserved.

Ad.7 Disclaimers: The performance Measure is not a clinical guideline and does not establish a standard of medical care, and has not been tested for all potential applications. THE MEASURE AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

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

Ad.8 Additional Information/Comments: Not applicable.