MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0037
Measure Title: Osteoporosis Testing in Older Women (OTO)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of women 65-85 years of age who report ever having received a bone density test to check for osteoporosis.
Developer Rationale: This measure assesses the number of women age 65-85 who report ever having received a bone density test to check for osteoporosis. There is convincing evidence that bone mineral density tests in women 65 years of age and older predicts short-term risk for osteoporotic fractures. There is also evidence that osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and in some cases, lead to mortality.

Numerator Statement: The number of women who report having ever received a bone mineral density test of the hip or spine.
Denominator Statement: Women age 65-85.
Denominator Exclusions: Women who received hospice care during the year.

Measure Type: Process
Data Source: Instrument-Based Data
Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Dec 30, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- The developer briefly described the link between bone mineral density and the patient's health outcomes in reduced risk of developing osteoporosis or sustaining a fragility fracture and reduced risk of morbidity and mortality.
- The developer provided a draft US Preventive Services Task Force Recommendation (release April 9,2018) including recommendations for the following:
  - "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool."
  - Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
  - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the draft US Preventive Services Task Force Recommendation (2018).

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided an updated (although still Draft) US Preventive Services Task Force Recommendation (released April 9, 2018) which continues to support their measure focus.

Exception to evidence

NA

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*
  - *If derived from patient report, does the target population value the measured process or structure and find it meaningful?*

Guidance from the Evidence Algorithm

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

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

RATIONALE:

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

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

Performance Data:

- Developer provided performance data for Medicare from HEDIS data from 2013, 2014, and 2015. The mean ranged from 74.4% to 75%

Disparities:

- Developer did not provide disparities from the measure. However cited a national cohort study by Gillespie and Morin that examined claims data from 2008 to 2014 for trends in osteoporosis screening in women age 50 and older. The data was categorized based on race/ethnicity, age, sex, and socioeconomic status.
    - They found that after controlling for other factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001).
    - After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening.
    - Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis (Gillespie and Morin).
- In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015).

Questions for the Committee:

- *Specific questions on information provided for gap in care.*
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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)

Evidence

- Evidence is rated as fair to good quality based in numerous studies and the USPSTF that screening identifies pts at risk and that treatment subsequently reduces fracture in post-menopausal women. 60 to 86.5 % of eligible patients performed screening, which indicates that the population for whom this is intended considers the process important.
- The evidence has been updated and is stronger than the previous review. There is no need to repeat the discussion and to vote on the evidence.
- Evidence from USPSTF has been updated in 2018 on the process measure for Osteoporosis Testing in Older Women (OTO). There is still no good or fair evidence for ""reducing fractures and fracture-related morbidity and mortality in adults."" However the USPSTF has graded the recommendation a ""B"" which means there is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
    - Committee may want to discuss the question: Is the evidence directly applicable to the process of care being measured?
- I am not aware of evidence or studies other than what was provided by the measure Developer. The Developer provided the following:
    - The developer provided an updated (although still Draft) US Preventive Services Task Force Recommendation (released April 9, 2018) which continues to support their measure focus.
- The measure focus is to determine what percentage of women enrolled in a Medicare Advantage Plan who are surveyed by a paper questionnaire have ever had a bone density test. Osteoporosis is a serious health problem and risk of fracture due to osteoporosis can be determined by bone density testing. Treatment is available to lessen the risk of fracture. The evidence for this Measure has been updated and has been rated "Moderate"
- This is a process measure of the percentage of women age 65-85 who have ever received a bone mineral density test of the spine or hip [not otherwise specified], excluding only those who have received hospice care during the year. This is measured by HEDIS data. Osteoporosis Testing in Older Women continues to be a HEDIS measure in 2018. The additional evidence is the draft USPSTF recommendation (4/9/18, Grade B) for DEXA

screening for women at age 65 and older or earlier if risk factors are present. The relationship between DEXA scores in the osteoporotic range and the risk of significant fractures is moderate.

- Reports of a survey of women who are at risk for osteoporosis, regarding bone mineral density testing. Although self reported, appropriate for this process measure.
- There is one reference to patient/consumer input but it is not descriptive and relates to the appropriateness of screening which is a different measure – this is a self-report collected through what is potentially an expensive process. Did patient find their self-awareness of this information to be meaningful? What is the societal cost per question in a survey of this scope – what questions are not asked?
- There is no evidence to support that self-awareness of osteoporosis screening adds value to patients – does this particular piece of health information equate to higher rates of activation?
- It appears the whole submission is aligned with a screening which this is not – unsure how to evaluate this.
- Developer provided systematic review and updated evidence (USPSTF April 2018) and graded quality of evidence supporting the rationale: that bone mineral testing predicts risk of fracture, that treatment reduces risk of fracture and fracture can lead to poor health, loss of function and death;
- No need for repeat discussion and vote on evidence
- Is there relationship to outcomes? Yes, see rationale above
- Strength of the evidence: moderate
- Is evidence applicable to process being measured? Yes
- Rating for evidence: moderate per QQC (box 4)
- The evidence provided is somewhat tangential. While there is a documented link between bone mineral density and health outcomes, there is no specific link between the act of measuring bone mineral density and health outcomes.

Performance Gaps

- There is performance gap, in that rates of screening increased from 2013 to 2014 then the lower percentile of performance decreased in 2015.
- Data included disparities in care between Caucasian, Afro-American, Asian and Hispanic women in performance of screening by age group
- There is a performance gap that exists. There seems to be some modest temporal improvement.
- No direct measure of performance gap, but other citations regarding the significantly lower screening rates for Black women and for women with lower socio-economic status were reported by the Developer.
- Committee may want to discuss question: Is there a gap in care that warrants a national performance measure?
- "Developer provided performance data for Medicare from HEDIS data from 2013, 2014, and 2015. The mean ranged from 74.4% to 75%. There is still an opportunity for improvement in this care category. No data on subgroups was provided. The Developer did cite the following studies on disparities:
    - Developer did not provide disparities from the measure. However cited a national cohort study by Gillespie and Morin that examined claims data from 2008 to 2014 for trends in osteoporosis screening in women age 50 and older. The data was categorized based on race/ethnicity, age, sex, and socioeconomic status.
    - They found that after controlling for other factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001).
    - After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening.
    - Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis (Gillespie and Morin).
    - In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015).
- HEDIS data for Medicare Advantage Programs showed variation in the rate of women who said that they received a bone mineral density test at some point in their life. In 2015, there was a 26.5 percentage point difference between Medicare plans at the 10th percentile and plans at the 90th percentile. This shows need for improvement. Ethnic, racial differences and socioecoonomic issues were not addressed. The developer did cite another report indicating that disparities affected access to testing.
- The developer provided performance data for Medicare from HEDIS data from 2013, 2014, and 2015. The mean ranged from 74.4% to 75%. HEDIS data from 2016 and 2017 were not provided by the developer. Medical literature published since 2014 has some studies indicating disparities in DEXA scanning by race and socioeconomic status exist.

- There remains a gap between the 10 and 90%ile for this measure over 25% - This is a gap that can be improved to improve the health of this population. No measure of ethnic, racial or social/economic differences, that may exist as described in the literature presented by the NCQA.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- Developer cites evidence which demonstrates racial and socioeconomic disparities in bone mineral testing;
- Gap in care that warrants national performance measure? Yes
- Opportunity for Improvement: moderate
- There are gaps in measurement of bone mineral density noted in the literature. However, this does not appear to be the appropriate tool to address this. Addressing these gaps warrants a national performance measure- but the use of other measures would be more impactful. Disparities were discussed, but there was no discussion is disparities have been identified through the use of this measure. It was not noted if the extremely small focus groups utilized to determine the survey language used for this measure were diverse in terms of race/ethnicity, socioeconomic status, or health literacy.

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.

Composite measures only:

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

NA

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

Evaluators: Primary Care and Chronic Illness project team staff

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

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

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

Committee pre-evaluation comments

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

Reliability Specification

- No concerns
- No concerns about consistent implementation.
- Data specifications are clearly defined however, it is important to note that the data is collected using a survey method.
- As before, concern exists over reliability of the respondent to questionnaire, either the patient completing the survey or her proxy.
- Patient report.
- Concur with the analysis of the staff evaluator.
- Reliability: no concerns that measure can be consistently implemented (measure specs are adequate)
- Reliability: no need to discuss and to vote on reliability
- Reliability rating: moderate
- The data elements are clearly defined for the develops but perhaps not for the patients who are assessed in this survey.

Reliablity Testing

- No concerns
- There are no concerns related to reliability.
- No concerns about Reliability.
- The data has been consistently collected year over year. No concerns with the reliability of the measure, however the data is dependent on patient reporting.
- Assuming consistency in this HEDIS measure, moderate reliability as judged in the initial submission of this measure should continue.
- No concerns
- Concur with the analysis of the staff evaluator.
- No concerns
- This measure is fatally flawed in that it relies upon patients (or their surrogates) to remember if they have had bone mineral density testing, to know what bone mineral density testing is, and to understand what osteoporosis or "brittle bones". It may be even less likely that their surrogate would know this information. While the results of this survey would be an interesting research project, the data provided, based solely on patient understanding and recollection, are not precise enough to determine or compare levels of performance.

Validity

- No concerns, but the developers on page 38 refer to "Osteoporosis management in women who had a fracture" rather than screening in women aged 65 to 85.
- No concerns about validity.
- This is an older population. As surveys move towards electronic collection there could be a reduction in number of respondents and may bias based on computer literacy.
- No concerns regarding validity testing results.
- Assuming consistency in this HEDIS measure, moderate validity as judged in the initial submission of this measure should continue.
- No validity concerns other than patient report
- Concur with the analysis of the staff evaluator.
- Validity: no concerns with validitity
- As noted in my response regarding reliability, the results from the survey described in the measure are

not reliable and are not necessarily valid in measuring quality of care.

Other Threats to Validity

- Patients who refuse testing, did not answer surveys, are illiterate or do not understand the survey questions
- The measure developer has gone through many steps to insure validity
- It appears that the measure targets the correct population, post-menopausal women
- Risk variables are all present at the start of care
- This measure was not risk adjusted
- There are no significant threats to validity.
- Developer reported on exclusions, which seemed not to suggest the exclusions was causing an threat to validity.
- No concerns with validity. Threat are as mentioned above, access to and ability to use computers in the elderly as surveys move towards electronic methods.
- No threats to validity from either exclusions or risk adjustment.
- Concur with the analysis of the staff evaluator.
- exclusions – Hospice, which based on .7% excluded is reasonable
- N/A

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

Data Specifications and Elements

- The measure has Information is gathered through the Health Outcome Survey (HOS). The Health Outcomes Survey is conducted through mailed surveys with telephone follow-up. (Per developer, there is concern that some/many Medicare beneficiaries do not have access to a computer or internet to complete the survey in electronic format. There is also a concern that moving to an internet-based mode of administration will bias results, as older frail adults may be less likely to complete the survey using an internet mode.)
- No data elements are in defined fields in electronic sources.
- Developer shared no difficulties on the use of this measure.
- This is not an eMeasure

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- This data was collected using the Health Outcomes Survey via mailings, return mail an f/u phone calls. it seems there are easier and more accurate electronic billing/administrative data that could accurately report the numerator and denominator.
- The data is collected both from administrative claims as well as EHR data. Both are byproducts of routine care delivery.
- Part of the Health Outcomes Survey (HOS) and no concerns about feasibility for data collection.
- Uses a survey method to collect the data. Data not collected during care delivery. The measure has Information is gathered through the Health Outcome Survey (HOS). The Health Outcomes Survey is conducted through mailed surveys with telephone follow-up. (Per developer, there is concern that some/many Medicare beneficiaries do not have access to a computer or internet to complete the survey in electronic format. There is

also a concern that moving to an internet-based mode of administration will bias results, as older frail adults may be less likely to complete the survey using an internet mode.)

- No data elements are in defined fields in electronic sources.
- Developer shared no difficulties on the use of this measure.
- This is not an eMeasure
- This measure is addressed by a paper survey developed by the NCQA and sent to the patient.
- The initial application mentions that the data for this measure are dependent on the Health Outcomes Survey (HOS), which is gathered by mailed surveys to patients covered by telephone contact. This particular measure is following the percentage of women screened by bone density measurements, which is a process and not an outcome measure. It is unclear if HOS data is needed or if the computation of the percentage of women in the target population screened can be gathered by provider data alone.
- This is a patient reported survey, paper format. If in the future other mechanisms to collect data, especially web-based data collection, this population may not respond in the manner currently using the paper format.
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- Required data elements are routinely captured during care delivery
- Potentially available in EHR or HIO
- Yes, mostly via mail (concerned with elderly not being able to access computer/internet)
- Feasibility rating: moderate
- Data are obtained from a survey, not EHR. The results generated are reliant upon the understanding and memory of a patient (or their surrogate). The focus groups utilized to develop this measure were too small to identify the potential impacts of health literacy. The data from this are not reliable or valid enough to measure quality of care or to effect change in practices.

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
NCQA STATE OF HEALTH CARE QUALITY ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported by WedMD and on the NCQA website. These ratings are based on a plan''s performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans

and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.

MEDICARE ADVANTAGE DISPLAY PAGE: This measure is listed on the display page for Medicare Advantage (Medicare Part C). This means that while performance on this measure is not tied to incentives; plans have the option to report.

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
This measure uses the following methods to obtain input: including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.

Questions received through NCQA's Policy Clarification Support system and above methods have informed how the developer revises the measure. However, the developer noted that health plans have not reported significant barriers to implementing the measure as it is collected through the Medicare Health Outcome Survery.

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

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

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

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

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

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

Improvement results
- The developer states that current HEDIS rates indicate that just under three quarters of women over the age of 65 in Medicare Advantage plans report having received at least one bone mineral density test in their lifetime. In 2015, the spread in national health plan performance was 60.0 to 86.5 percent (10th to 90th percentiles).

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
- Per developer, there were no identified unexpected findings (positive or negative) during testing or since implementation of this measure.

Potential harms

- The developer did not identify any potential harms in testing. Per the evidence form by developer, developer noted the following by USPSTF: "The USPSTF found no studies that described harms of screening for osteoporosis in men or women. Based on the nature of screening with bone measurement tests and the low likelihood of serious harms, the USPSTF found adequate evidence to bound these harms as no greater than small. Harms associated with screening may include radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)."

Additional Feedback:

- In 2015 NQF Endocrine Report, the Committee mentioned concern about proxy and/or patients with cognitive impairment answering the survey.

***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

Use

- It is used in various public reporting measures, such as NCQA state of health care quality, Medicare advantage plan reporting, and health plan ratings
- NCQA reports their data via conferences and webinars. They also provide performance benchmarks to permit health plans to gauge their success.
- They have received feedback from multiple stakeholders and advisory panels as well as public commenting.
- This data is currently used in the CMS QPP program.
- Accountability and Transparency are well documented. Feedback on measure is included.
- Committee may want to discuss question: How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?"
- NCQA STATE OF HEALTH CARE QUALITY ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population.
- NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported by WedMD and on the NCQA website. These ratings are based on a plan''s performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.
- MEDICARE ADVANTAGE DISPLAY PAGE: This measure is listed on the display page for Medicare Advantage (Medicare Part C). This means that while performance on this measure is not tied to incentives; plans have the option to report."
- Accountability and feedback are provided by NCQA using HEDIS data
- HEDIS reports are available to health plans, clinicians, and other health care organizations, but apparently in aggregate without information on individual patients being available in this system. Given stability in HEDIS measures, the "report card" value of comparing percentages of applicable women screened comparing year to year performance would be useful.
- This data is reported to health plans for action, and is part of their evaluations at national level.
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity.
- Measure is currently publically reported and used in accountability program
- Ratings from 521 Medicare Advantage plans, 614 commercial plans, 294 Medicaid plans

- Developer states no reported barriers to implementation from health plans as measure is collected via Medicare Health outcome Survey
- How have results been used to further goal of high quality care? Results reported on NCQA State of Health Care Annual Report, the NCQA Health Plan rating report cards and on NCQA website and reported on Web-MD
- Usability Rating: Pass
- The results of the measure have been publicly reported. Other than reporting results, there is no indication how this data has been utilized to influence practices or health plans.

Usability

- Overall, about 75% of women aged 65 to 85 had screening. The 10% percentile was about 60%, the 90th percentile was 86.5%. page 24 shows that the latest reported data, from 2015 sowed a lower 10th percentile of patients accomplishing screening than the prior to years by about 1-2%, but a higher 90% percentile accomplishing from .3 to .9 %
- No unintended consequences of screening
- There are not evident unintended consequences.
- Benefits and harms seemed to be the one area where there is a dearth of studies to assist in evaluating the screening for osteoporosis as part of the evaluation. However, all parties concur that without such studies, the harms appear to be small and are outweighed by the benefits.
- No identified harms from this measure. No unintended consequences, other than bias that may result if only use electronic means to collect survey results.
- Efforts should be made to increase the percentage of this performance measure so that more older women at risk for fracture are tested.
- No harm from this measure.
- As screening with DEXA scans in this population is suggested by the USPSTF at a Grade B level, securing and following the percentage of the target population so screened would be a benefit to the providers and to the population served by them. A complicating factor is the relatively recent decrease in health insurance coverage for DEXA scans, which appears to be significant in reducing the number of women in this age group receiving screening.
- No concerns. Benefits outweigh harms. There is need for improvement in this measure.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- 75% of women 65-85 have received at least 1 BMT; can reduce disparities among racial and socioeconomic lines;
- benefits outweigh any potential risks (none identified)
- No appreciable harms noted by developer (?radiation exposure, opportunity cost)
- Usability rating: moderate
- There are benefits to obtaining bone mineral density testing and treating low bone mass, when identified. While there are complications (e.g., atypical femur fractures) from the use of many of the medications used to treat low bone mass, these risks outweigh the benefits. However, if we are going to assess quality of care through assessing for and treating low bone mass, then health care providers and health plans should be assessed based on ordering these test, assuring the tests are completed, and acting on the results. While obtaining this information from a patient's health record may be difficult, as the DXA may have been performed prior to the patient initiating care at a given practice, patient care would be more likely to be impacted by working through this issue, rather than obtaining their responses to a survey such as described in this measure.

Criterion 5: Related and Competing Measures

Related or competing measures

- Developer identified three related or competing measures
  - 0046 : Screening for Osteoporosis for Women 65-85 Years of Age
  - 0053 : Osteoporosis Management in Women Who Had a Fracture
  - 2417 : Risk Assessment/Treatment After Fracture

Harmonization

- Related/Competing Measures:

- Measure 0046 assesses the percentage of women who have a bone mineral density test to screen for osteoporosis. Measure 0046 is collected using medical record review and is only specified for physician level reporting. The rationale for different data sources is the availability of data for the level of reporting. The developer describes in further detail the harmonization of these two measures in/since 2014. Both measures have same steward (NCQA). (Competing)
- Measure 0053 addresses a different population than 0037 (i.e., women who have experienced a fragility fracture), and is therefore focused on secondary prevention of future fractures as opposed to screening for osteoporosis. (Related)
- Measure 2417 also focuses on those who had a fragility fracture and then received secondary prevention. (Related)

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

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0037
Measure Title: Osteoporosis Testing in Older Women

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

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

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

   ☒Yes (go to Question #2)

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

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?
REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

4. 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.*
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☒No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)
☒No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
   ☒Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:
    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)
    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?
REFERENCE: Testing attachment, section 2b2.
***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*
☐Yes (please explain below then go to Question #13)
☒No (go to Question #13)
☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?
***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
☐Yes (please explain below then go to Question #14)
☐No (go to Question #14)
☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.
☐Yes (please explain below then go to Question #15)
☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
    ☒Yes (go to Question #18)
    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
    ☒Yes (go to Question #19)
    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
    ☒Yes (go to Question #20)
    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #21)
    ☒Moderate (go to Question #21)
    ☐Low (please explain below then go to Question #21)
    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Prior validity studies of the same data elements may be submitted*
    ☒Yes (go to Question #22)
    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*
    ***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.
***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

REFERENCE: Testing attachment, section 2c

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

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0037

Measure Title: Osteoporosis Testing in Older Women

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

Date of Submission: 4/9/2018

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: Click here to name the health outcome

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

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Screening for Osteoporosis

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

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

2014 Submission

Female patients at risk for osteoporosis (age 65 and older)>>> bone mineral density test to check for low bone mass or osteoporosis >>> low bone mass identified >>> patient evaluated for treatment options >>> treatment >>> reduced risk of developing osteoporosis or sustaining a fragility fracture >>> maintained quality of life and reduced risk of morbidity and mortality.

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.)

2018 Submission

During the measure's reevaluation we sought feedback on the value and importance of the measure from our measurement advisory panels and through public comment. Patient and consumer representatives on our panels indicated that osteoporosis screening is a valued service for older women.

****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

USPSTF Recommendation: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission NCQA acknowledges that as of April 9, 2018, the U.S. Preventive Services Task Force (USPSTF) has released a DRAFT recommendation statement for osteoporosis screening. A draft Evidence Review was also published in November 2017. When published, NCQA will evaluate the final recommendation statement and supporting evidence review and consider any potential changes that may be needed for this measure. However, based on the draft recommendation statement we do not anticipate that any major revisions will be needed. U.S. Preventive Services Task Force. 2017. *Draft Recommendation Statement: Osteoporosis to Prevent Fractures: Screening*. https://www.uspreventiveservicestaskforce.org/Page/Document/draft-recommendation-statement/osteoporosis-screening1 U.S. Preventive Services Task Force. 2017. *Draft Evidence Review: Osteoporosis to Prevent Fractures: Screening.* https://www.uspreventiveservicestaskforce.org/Page/Document/draft-evidence-review/osteoporosis-screening1 2014 Submission U.S. Preventive Services Task Force. 2011. Screening for osteoporosis: US preventive services task force recommendation statement. Annals of internal medicine, 154(5), 356.

	http://www.uspreventiveservicestaskforce.org/uspstf10/osteoporosis/osteors.htm, accessed May 2, 2014.
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.	2018 Submission "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool." 2014 Submission "The USPSTF recommends screening for osteoporosis in women aged 65 years or older and in younger women whose fracture risk is equal to or greater than that of a 65-year-old white woman who has no additional risk factors."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission The USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in women age 65 years and older is at least moderate.
Provide all other grades and definitions from the evidence grading system	2018 Submission N/A
Grade assigned to the **recommendation** with definition of the grade	2018 Submission Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. 2014 Submission This measure is based on a grade B recommendation from the USPSTF. Grade B: The USPSTF recommends the services. 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	2018 Submission Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. 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. Grade I: 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. 2014 Submission Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. 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?	2018 Submission The DRAFT evidence report (Viswanathan et al 2017) supporting this guideline outlines the quantity and quality of evidence, which are summarized below for the key questions of the 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? • As in the previous 2011 review, found no good or fair quality studies eligible for this key question Key Question 2a. What is the accuracy and reliability of screening approaches to identify adults who are at increased risk for osteoporotic fracture? • Accuracy of Clinical Risk Assessment Tools for Identifying Osteoporosis: included 37 articles (35 studies, fair or good quality) • Accuracy of Bone Measurement Tests Used to Identify Low Bone Mass and Osteoporosis: included 11 studies, fair or good quality • Accuracy of Bone Measurement Tests Used to Predict Fracture: included 21 studies, fair or good quality • Accuracy of Fracture Risk Prediction Instruments: included 1 systematic review and 13 fair or good quality observational studies Key Question 2b. What is the evidence to determine screening intervals and how do these vary by baseline fracture risk? • Included 2 articles (2 studies, good quality) Key Question 3. What are the harms of screening for osteoporotic fracture risk? • Found no eligible studies that addressed this question Key Question 4a. What is the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality? Bisphosphonates: • Alendronate: included 7 studies, fair or good quality

- Zoledronic Acid: included 2 studies, fair or good quality
- Risedronate: included 4 studies, fair or good quality
- Etidronate: included 2 fair quality studies
- Ibandronate: identified no studies or trials that assessed the benefits of ibandronate for preventing fractures

Raloxifene:
- Included 1 large good quality RCT

Estrogen:
- No studies included

Denosumab:
- Included 3 fair quality trials

Parathyroid Hormone:
- Included 2 fair quality trials

Key Question 4b. How does the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality vary by subgroup, specifically in postmenopausal women, premenopausal women, men, younger age groups (age <65 years), older age groups (age ≥65 years), baseline bone mineral density, and baseline fracture risk?

Bisphosphonates:
- Zoledronic Acid, Etidronate, Ibandronate: found no relevant results in included studies for subgroup analysis
- Alendronate: included 1 study
- Risedronate: included 1 RCT

Raloxifene:
- Included 1 study

Estrogen:
- No studies included

Denosumab:
- Included 1 fair quality trial

Parathyroid Hormone:
- Included 1 fair quality trial

Key Question 5. What are the harms associated with pharmacotherapy?

Bisphosphonates:
- Alendronate: included 16 studies, fair or good quality
- Zoledronic Acid: included 4 studies, fair or good quality
- Risedronate: included 4 studies, fair or good quality
- Etidronate: included 2 fair quality studies
- Ibandronate: included 7 fair quality studies

Raloxifene:
- Included 6 studies

Estrogen:
- No studies included

Denosumab:
- Included 3 fair quality studies

Parathyroid Hormone:
- Included 2 fair quality studies

Viswanathan, M., et al. 2017. "Screening to Prevent Osteoporotic Fractures: An Evidence Review for the U.S. Preventive Services Task Force." Available here:

	https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryDraft/osteoporosis-screening1 **2014 Submission** Quantity: N/A (not required for previous submission) Quality: N/A (not required for previous submission)
Estimates of benefit and consistency across studies	2018 Submission The following text is quoted directly from the USPSTF recommendation statement. The USPSTF found no studies that evaluated the effect of screening for osteoporosis on fracture rates or fracture-related morbidity or mortality. The USPSTF found convincing evidence that bone measurement tests are accurate for detecting osteoporosis and predicting osteoporotic fractures in women and men. The USPSTF found adequate evidence that clinical risk assessment tools are moderately accurate in identifying risk of osteoporosis and osteoporotic fractures. The USPSTF found convincing evidence that drug therapies reduce subsequent fracture rates in postmenopausal women. The benefit of treating screening-detected osteoporosis is at least moderate in women age 65 years and older and younger postmenopausal women who have similar fracture risk. The harms of treatment range from no greater than small for bisphosphonates and parathyroid hormone to small to moderate for raloxifene and estrogen. Therefore, the USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in these groups of women is at least moderate. The USPSTF concludes that the evidence is inadequate to assess the effectiveness of drug therapies in reducing subsequent fracture rates in men without previous fractures. Treatments that have been proven effective in women cannot necessarily be presumed to have similar effectiveness in men, and the direct evidence is too limited to draw definitive conclusions. Thus, the USPSTF could not assess the balance of benefits and harms of screening for osteoporosis in men. **2014 Submission** N/A
What harms were identified?	2018 Submission The following is quoted directly from the USPSTF draft recommendation statement: "The USPSTF found no studies that described harms of screening for osteoporosis in men or women. Based on the nature of screening with bone measurement tests and the low likelihood of serious harms, the USPSTF found adequate evidence to bound these harms as no greater than small. Harms associated with screening may include radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)."

	2014 Submission N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	2018 Submission To our knowledge, there have been no published studies since the systematic review that would impact the recommendations. **2014 Submission** N/A

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0037
Corresponding Measures:
De.2. **Measure Title:** Osteoporosis Testing in Older Women (OTO)
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of women 65-85 years of age who report ever having received a bone density test to check for osteoporosis.
1b.1. **Developer Rationale:** This measure assesses the number of women age 65-85 who report ever having received a bone density test to check for osteoporosis. There is convincing evidence that bone mineral density tests in women 65 years of age and older predicts short-term risk for osteoporotic fractures. There is also evidence that osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and in some cases, lead to mortality.

S.4. **Numerator Statement:** The number of women who report having ever received a bone mineral density test of the hip or spine.
S.6. **Denominator Statement:** Women age 65-85.
S.8. **Denominator Exclusions:** Women who received hospice care during the year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Instrument-Based Data
S.20. **Level of Analysis:** Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Dec 30, 2014

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

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0037_OTO_Evidence.docx
1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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 assesses the number of women age 65-85 who report ever having received a bone density test to check for osteoporosis. There is convincing evidence that bone mineral density tests in women 65 years of age and older predicts short-term risk for osteoporotic fractures. There is also evidence that osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and in some cases, lead to mortality.

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.*
PERFORMANCE RATES: The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data is summarized at the health plan level and described by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year.

The data below demonstrates the variation in the rate of women who said that they received a bone mineral density test at some point in their life. In 2015, there was a 26.5 percentage point difference between Medicare plans at the 10th percentile and plans at the 90th percentile. This gap in performance underscores the opportunity for improvement.

Medicare Performance
Osteoporosis testing among all women ages 65 and older
YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH
2013 | 74.8% | 9.3% | 61.5% | 68.9% | 76.8% | 81.9% | 85.6%
2014 | 75.0% | 9.5% | 61.7% | 69.3% | 76.6% | 82.1% | 86.2%
2015 | 74.4% | 9.9% | 60.0% | 67.7% | 76.0% | 81.8% | 86.5%

The data shown above are from HEDIS data collection reflecting the most recent years of data for this measure. In 2016, HEDIS measures covered 17.6 million Medicare members from 495 Medicare Advantage Organizations. The rate for each plan is collected from the Health Outcome Survey; in 2016 the response rate for the survey across 463 plans that fielded the survey was 45 percent, resulting in 302,404 completed surveys. The number of health plans reporting, response rate, and number of completed surveys was similar across years.

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.
N/A

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.*
NCQA does not currently report performance data stratified by race, ethnicity. While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity collected from the survey.

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
In a national cohort study (Gillespie and Morin 2017), researchers examined medical claims data from 2008 to 2014 for trends in

osteoporosis screening in women age 50 and older. They found that after controlling for other factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001). After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening. Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis (Gillespie and Morin). In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015). Interventions that target population screening are needed to improve the rates of osteoporosis screening for all women age 65 and older, but particularly for Black women and those with lower socioeconomic status.

Amarnath, A. L. D., Franks, P., Robbins, J. A., Xing, G., & Fenton, J. J. (2015). Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study. Journal of general internal medicine, 30(12), 1733-1740.
Gillespie, C. W., & Morin, P. E. (2017). Trends and disparities in osteoporosis screening among women in the United States, 2008-2014. The American journal of medicine, 130(3), 306-316.

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):*
Endocrine, Musculoskeletal, Musculoskeletal : Osteoporosis

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

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

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.)*
www.hosonline.org

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

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.
Attachment **Attachment:** OTO_spec_hos_hedis_volume6_2018.pdf

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.
Patient

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Added an exclusion for patients receiving hospice care.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The number of women who report having ever received a bone mineral density test of the hip or spine.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The number of female patients 65-85 years of age who responded "yes" to question 52 in the Medicare Health Outcomes Survey.

Question 52: "Have you ever had a bone density test to check for osteoporosis, sometimes thought of as 'brittle bones'? This test would have been done to your back or hip."

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women age 65-85.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The number of women 65-85 years of age who responded to question 52 on the Medicare Health Outcome Survey.
Question 52: "Have you ever had a bone density test to check for osteoporosis, sometimes thought of as 'brittle bones'? This test would have been done to your back or hip."

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Women who received hospice care during the year.

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

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

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 = Higher 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.)*
Step 1: Identify the eligible population – Of those who were selected to receive a survey, identify all female patients age 65-85 who answered Question 52: "Have you ever had a bone density test to check for osteoporosis, sometimes thought of as 'brittle bones'? This test would have been done to your back or hip."
Step 2: Determine the number of patients in the eligible population who responded "Yes".
Step 3: Calculate a rate (the number of patients who responded "yes" divided by the eligible population)

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.
Sampling: This measure is collected through the Medicare Health Outcomes Survey: a 64 item mailed survey with telephone follow up. This survey is conducted by certified survey vendors to health plan beneficiaries in their home or place or residence. To allow for adequate sample size, within a health plan, a random sample of 1,200 beneficiaries is surveyed (if a health plan has fewer than 1,200 members, all members of the health plan are sampled). Organizations with fewer than 500 members are excluded from sampling.

Proxy responses: The Health Outcome Survey allows for a family member or "proxy" to fill out the survey. The survey is mailed to patients with the following instructions: "If you are unable to complete this survey, a family member or "proxy" can fill out the survey about you." At the end of the survey, the respondent is asked the following question:

Who completed the survey form?
Answer= "Person to whom survey was addressed" or "Family member or relative of person to whom the survey was addressed" or "Friend of person to whom the survey was addressed" or "Professional caregiver of person to whom the survey was addressed."

This information is used to determine if information from proxy respondents is systematically biased or different from patient self-reported data.

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.
The standard protocol for administering the Medicare Health Outcomes Survey employs a combination of mail and telephone administration. The main data collection technique is a mailing of surveys to sampled members. If members fail to respond after two mailings, survey vendors attempt at least six telephone follow-up calls. In addition, if members return a blank or incomplete mail survey, survey vendors attempt at least six telephone follow-up calls to obtain response to unanswered questions. NCQA does not allow the organization or survey vendor to use incentives of any kind.

Minimum Response Rate: To ensure reliable comparisons between health plans a minimum sample size of 100 in the measure denominator is required.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Instrument-Based Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
The Medicare Health Outcome Survey can be administered by mail or telephone using a CATI protocol. It is offered in English, Spanish, and Chinese (mailed survey only). Detailed instructions for the administration of the Health Outcomes Survey and the complete survey can be found at www.hosonline.org.

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)*
Available at measure-specific web page URL identified in S.1

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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

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

2. Validity – See attached Measure Testing Submission Form
0037-Testing_Form_v7.1_-636555245641753088.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0037

Measure Title: Osteoporosis Testing in Older Women

Date of Submission: 4/9/2018

Type of Measure:

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

Instructions

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

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

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

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

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator

exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record

☐ claims	☐ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Patient Reported Data/Survey	☒ other: Patient Reported Data/Survey

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?
2018 Submission:
Sample 3: 2016

2014 submission:
Sample 1: 2005
Sample 2: 2012

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

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

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
2018 Submission:
Sample 3: To test the incidence of the exclusion for individuals receiving hospice care (first incorporated into the measures in 2016), we used data from Medicare health plans submitting Health Outcome Survey data to be reported in HEDIS for measurement year 2016. The plans were nationally representative and included 463 PPO and HMO plans.

2014 submission:
Sample 1: To test data element reliability and validity, NCQA contracted with RTI to conduct four rounds of cognitive testing between January and May 2005 in Raleigh and Durham, North Carolina, and Waltham, Massachusetts. Six respondents in each round for a total of 24 completed interviews. There were two rounds of concept testing to identify which terms used to describe osteoporosis and osteoporosis testing were recognized and understood by respondents. Using the terms identified in Rounds 1 and 2, the osteoporosis survey question was then tested in Rounds 3 and 4.

Sample 2: This measure was tested for reliability, empirical validity, meaningful difference in performance and missing data using data from Medicare health plans submitting Health Outcome Survey data to be reported in HEDIS for measurement year 2012. The plans were nationally representative and included 495 PPO and HMO plans.

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*)

2018 Submission:

Sample 3: In 2016, this measure was collected from 302,404 survey responders (patients) from 463 health plans. A sample was drawn from each health plan's population based on the plan size (500 to 1,200 patients per health plan are sampled depending on plan size).

2014 submission:

Sample 1: Four rounds of cognitive testing took place between January and May 2005. Six respondents were interviewed in each round for a total of 24 completed interviews. The first two rounds of testing were done to identify which terms used to describe osteoporosis and osteoporosis testing were recognized and understood by respondents. The survey question was then tested in rounds three and four and were based off of the terms that were identified in rounds one and two. Participants were recruited for each round of cognitive testing from senior centers, senior housing, physical therapy, wellness centers, physicians' offices and by word of mouth. In addition, announcements were placed about the study in local newspapers. Respondents were also required to have seen a health care provider during the past year. For the fourth round of testing, women aged 65 and older who had been diagnosed with osteoporosis were recruited. Round four included two women 65-75 years of age and four women who were over the age of 75. Respondents in round four were also diverse on their level of education. One respondent had less than high school, one had some high school, three were high school graduates or had their GED, and one woman had a four-year college degree or more.

Sample 2: In 2012, this measure was collected from 297,974 survey responders (patients) from 495 health plans. A sample was drawn from each health plan's population (1,200 beneficiaries per health plan sampled). Table 1 below lists the demographic characteristics of the 2012 cohort.

Table 1: Demographic Characteristics of Sample 2 (Health Outcome Survey 2012 Cohort)

		%
Age	Under 65	15.5
	65–69	25.4
	70–74	22.1
	75–79	16.3
	80 and older	20.8
Gender	Male	42.5
	Female	57.5
Race	Hispanic	3.2
	North American Native	0.3
	Asian	2.2
	Black	12.2
	White	79.6
	Other	2.1
	Unknown	0.4

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.

2018 Submission:

Sample 3 was used to test the incidence of the exclusion for individuals receiving hospice care.

2014 submission:

Sample 1 was used to test item-level reliability and validity.

Sample 2 was used to test reliability, empirical validity, meaningful difference in performance, and missing data.

Validity was also demonstrated through a systematic assessment of face validity. This measure was systematically evaluated for face validity with four panels of experts:

- The Osteoporosis Advisory Workgroup included 5 experts in geriatrics, endocrinology, and osteoporosis.
- The Geriatric MAP included 13 experts in geriatrics, including representation by consumers, health plans, health care providers and policy makers.
- The Technical Measurement Advisory Panel includes 14 members, including representation by health plans methodologists, clinicians and HEDIS auditors.
- NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 21 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

Per NQF instructions we have described the composition of the expert panels which assessed face validity for this measure. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panels.

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

2018 Submission:

We did not analyze performance by social risk factors.

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*)

2014 submission:

Reliability Testing of Performance Measure Score: In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient." This approach is also relevant to health plans and other accountable entities.

Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

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

2014 submission:
Results of Reliability Testing of Performance Measure Score:

Table 2: Reliability in Medicare Plans in 2012

# of plans	Overall Reliability Score	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile
495	0.995	0.994	0.995	0.996	0.997	0.997

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?*)

2014 submission:

Interpretation of measure score reliability testing: Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The reliability output from HEDIS 2012 data shows high overall reliability with a mean individual reliability above .9 for all plans. The lowest individual reliability found among plans was .92, the highest individual reliability was .99. Reliability assesses the degree to which a measure produces stable and consistent results, therefore, there is a high degree of consistency in the results and the variability between plans is most likely due to the performance of plans. Eight plans did not have a denominator of >30 and were not included in the reliability analysis.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

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

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

2014 submission:

Method of Testing Critical Data Element Validity: Cognitive testing is a process used routinely in determining the content validity of survey questions. Through cognitive testing, trained interviewers assess whether respondents understand survey questions, can recall the information being asked of them, and answer the questions correctly given their experiences. Cognitive interviewing involves asking volunteer respondents to answer the survey questions (either on paper or verbally) and then interviewing respondents about their answers to the questions. Interview protocols include specific questions about the respondent's thought-process when answering the questions. Questions are tested in rounds to allow for revision to the survey questions and interview protocol between testing rounds. Testing was completed by a professional research team at RTI. The text below describes the specific testing protocol in greater detail:

There were four rounds of testing for this measure. The first two rounds of testing focused on concept testing with the goal of determining whether women were familiar with the term "osteoporosis," the best term to use as a descriptor of osteoporosis, and the best term to use for osteoporosis testing. We also tested the effectiveness of adding a description of the osteoporosis test.

The following terms were tested in Rounds 1 and 2 (terms that did not test well were dropped for the second round of concept testing):

Osteoporosis (descriptors)

- Bone loss
- Weakening of the bones
- Thin bones
- Brittle bones

Osteoporosis testing

- Bone density test
- Bone scan
- Dexa-scan
- Densitometry
- Bone mineral test
- Bone ultrasound
- BMD test

2018 Submission:

Method of assessing face validity: We describe below NCQA's process for both measure development, and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web

Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assures that measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure "Desirable"? The work-up is vetted by NCQA's MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation

process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2.

Method of Testing Empirical Validity: We tested for construct validity by exploring whether performance for this measure was correlated with a similar measure, Osteoporosis Management in Women Who Had a Fracture. This measure assesses the percentage of women who experienced a fracture and received either bone mineral density test or a prescription for an osteoporosis treatment. The intent of the Osteoporosis management measure is to assess a health plan's performance at secondary prevention of osteoporosis related fracture. We specifically hypothesized that these two measures would be positively correlated (i.e. plans that have high rates of performance for management of osteoporosis will also have high rates of performance for screening of osteoporosis.) To test this correlation we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

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

2014 submission:

Results of Critical Data Element Validity Testing: Most women had heard the term "osteoporosis" before testing and were usually able to accurately describe it. Several descriptors were tested including "brittle bones," "bone loss," and weakening of the bones." Although not unanimous, most women picked "brittle bones" as their top choice of a descriptor, but even those who did not pick this term as a top choice were still familiar with it and thought it was an accurate way to describe osteoporosis. We tested a long list of terms used to describe the osteoporosis test. Most of the terms were technical and not known by respondents (e.g., DXA-scan, densitometry). The term "bone density test" was the term most familiar to respondents. The effectiveness of adding a description of the test was also examined ("This test may have been done to your back or hip"). All of the respondents found this addition to be helpful.

Results of Face Validity Assessment:

Step 1: This measure was developed in 2002 to address under-diagnosis and treatment of osteoporosis in women who had fragility fractures. NCQA, along with the Osteoporosis Technical Subgroup and the Geriatric Measurement Advisory Panel, worked together to assess the most appropriate screening and treatment for women who had a fragility fracture.

Step 2: The measure was written and field-tested in 2002. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2003.

Step 3: The measure was released for Public Comment in 2003 prior to publication in HEDIS. The CPM recommended moving this measure to first year data collection by a majority vote.

Step 4: The measure was introduced in HEDIS 2004. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure to public reporting with a majority vote.

Step 5: The measure was re-evaluated in 2013 and reviewed by the Osteoporosis Workgroup and the Geriatric Measurement Advisory Panel. The measure was presented to the CPM in January 2014 and proposed changes to the measure were posted for public comment February-March 2014. The CPM approved the proposed changes to the measure in May 2014 with a majority vote. These changes will go forward for use in HEDIS 2015.

Conclusion: The measure was deemed to have the desirable attributes of a HEDIS measure (relevance, scientific soundness, and feasibility).

Results of Construct Validity Testing: The results in Table 1a indicated that the Osteoporosis Testing measure was significantly (p<.05) correlated with the Osteoporosis Management measure (NQF #0053) in the direction that was hypothesized.

Table 3: Correlation between Osteoporosis Measures in Medicare Plans – 2012

Pearson Correlation Coefficient	
	Osteoporosis Testing in Older Women
Osteoporosis Management in Women who have had a Fracture	R=0.27305 (R Statistic) p<.0001 (significance)

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?*)

2014 submission:

Interpretation of Critical Data Element Validity Testing: Cognitive testing showed that the terms used in the measure are understandable and familiar to most women.

Interpretation of Construct Validity Testing: Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results confirmed the hypothesis that this measure is correlated with the Osteoporosis Testing in Older Women (NQF #0037), suggesting they represent the same underlying construct of quality of care for osteoporosis. Although the association was weak, it was significantly greater than zero. A strong correlation would not be expected in this case due to the different denominators of these two measures.

2018 Submission:

Interpretation of face validity assessment:

NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Osteoporosis Management in Women Who Had a Fracture* is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)

2018 Submission:

The exclusion for this measure for individuals in hospice care is based on using a Hospice Status flag in the file that contains all the CMS Health Outcome Survey data submission. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons. While we did not fully test this exclusion and its impact on measure performance, using data from measurement year 2016 we examined the total number and percent of individuals who were excluded from measure reporting based on the Hospice Status Flag.

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*)

2018 Submission:

The total Health Outcome Survey quality reporting sample was 668,143 individuals. Of these, 4,677 (0.7%) individuals had the Hospice Status flag and were excluded from the reporting of this measure.

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*)

2018 Submission:

The 0.7% incidence of those meeting the hospice exclusion criteria is in line with what we would expect to see given publicly available data published by CMS on the use of hospice services among Medicare Advantage beneficiaries. While we were not able to test the impact of the exclusion on performance measure score, excluding individuals in hospice care from getting bone mineral density tests to screen for osteoporosis makes clinical sense. The exclusion was implemented using data that was already collected and reported (the Hospice Status Flag) and therefore added no additional burden to measure reporting.

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

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

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

☒ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

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.

2014 submission:

This measure is a process measure collected from patient self-report. Although this measure is collected from patient self-report it is not a PRO-PM, as it does not assess a patient reported outcome. Therefore we do not risk-adjust the rates.

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?

2014 submission:

N/A

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?
2014 submission:
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.**
2014 submission:
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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
2014 submission:
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*)
2014 submission:
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)*
2014 submission:
To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the

difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities.

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*)

2014 submission:

Table 4: Variation in Performance across Medicare Health Plans in HEDIS (2012 data)

	Avg.	SD	10th	25th	50th	75th	90th	IQR
Medicare Plans	73.1	9.6	59.3	67.1	74.6	81.0	84.1	13.9

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range

Table 5: T-test between two randomly selected health plans in HEDIS (2012 data)

	Plan Rate (25th Percentile)	Plan Rate (75th Percentile)	Z-score	P-Value
Medicare	65.8	82.2	4.3	<.05

P-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

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?*)

2014 submission:

The results above indicate there is a 13.9 percent gap in performance between the 25th and 75th performing plans. The difference between the 25th and 75th percentile is statistically significant.

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

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

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

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

statistical analysis was used)

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2014 submission:

This measure is collected through the Health Outcomes Survey. Our analysis of missing data was done at two levels: survey-level and item-level.

Survey Level Missing Data Analysis: NCQA conducted an analysis in 2013 to investigate whether there were significant differences between responders, late responders and nonresponders to the Health Outcome Survey. Where data are collected in waves, such as with the HOS, there is the opportunity to estimate the potential impact of nonresponse by studying specific subsets of the responder pool. The classic concept of the continuum of resistance postulates that individuals who fail to respond to multiple survey attempts are increasingly more resistant to completing a survey and that the most difficult-to-reach respondents may be similar to nonrespondents (Halbesleen 2013). In the context of the HOS, members can be classified as "late responders" (time to survey completion exceeded the Cohort and administration-specific 90th percentile), "other responders" and "nonresponders." Because within the concepts of wave analysis, late responders may be more similar to nonresponders, the late responders population can be compared with all other responders to estimate how different true nonresponders may be from responders. We estimated differences in member characteristics across multiple years in a sample of responders and nonresponders (only 2010 data is displayed below, results were across multiple years). The characteristics compared across populations included: CMS-Hierarchical Condition Categories (HCC) risk score (an administrative proxy for health status), age, gender, race, disability status, hospice status, institutionalization status, and end stage renal disease (ESRD) status. These characteristics were available in CMS administrative systems and were not obtained through the Health Outcomes Survey. Additional characteristics for late responders and on-time responders were drawn from the HOS survey: household income, home ownership, marital status, and education level. Given the large sample size, differences between responders, late responders and nonresponders were evaluated using effect size calculations (Cohen's D for continuous variables and Cramer's V for nominal variables.) Cramer's V for nominal variables is used to examine the association of two values. The result is between 0 (no association) and +1 (complete association). Cohen's D calculates the difference between two means divided by the standard deviation for the data. This formula is typically used to estimate needed samples sizes although here it was used to compare differences in means between samples. A lower score indicates the need for a larger sample size where as a higher score indicates a direct correlation between two means.

Halbesleben, J.R.B., and M.V. Whitman. 2013. Evaluating Survey Quality in Health Services Research: A Decision Framework for Assessing Nonresponse Bias. Health Serv Res.Jun;48(3):913-30.

Item Level Missing Data Analysis: To further understand the potential impact of missing data we calculated the rate of item-level missing data for survey responders in 2012. We calculated the average rate of missing data on the osteoporosis question, the distribution of missing data across health plans and the frequency of missing data.

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

2014 submission:

Survey Level Missing Data Analysis:

In 2010 the response rate to the Health Outcome Survey was 63.3%, among responders 9.1% were "late responders" who responded (time to survey completion exceeded the Cohort and administration-specific 90th percentile.) Table 6 shows differences between responders, late responders and nonresponders in a sample from 2010, and the effect size of those differences. Due to sample size, nearly all differences in mean or proportion were significant, although most differences were small when evaluated in terms of effect sizes with the exception of race.

Table 6: Characteristics of Survey Responders, Late Responders and Nonresponders

	Mean or Percentage of Total			Significance and Effect Size[1]		
	Responder	Non	Late	Other vs. Non	Late vs. Other	Late vs. Non
Age[2,3]	72.9	71.4	71.5			
HCC[2,3]	1.1	1.2	1.1			
Male	42.9	44.8	43.4			
Non-White	19.0	27.5	23.5	**	*	
Disability	14.1	19.7	19.2	*		
Dual Eligibility	19.9	29.3	25.0	**		
Institutionalized	0.7	3.4	.4	**		*
Hospice	0.1	0.2	.1			
ESRD	0.1	0.1	.1			
Household Income (<20,000)	46.0	38.7		*		
Not a Homeowner	38.5	34.0				
Education Level—Less Than High School	37.5	27.6		*		
Marital Status—Not Married	45.9	47.2				

[1]Effect size estimates for pairwise comparison of nominal variables for responders and nonresponders were based on Cramer's V. The following classification was used: 0 to <0.05 = no effect; 0.05 to <0.1 = weak effect (*); 0.10 to <0.15 = moderate effect (**); 0.15 to <0.25 = strong effect (***); >0.25 = very strong effect (****).

[2]Effect size estimates for pairwise comparisons of continuous variable for responders and nonresponders were based on Cohen's D, whereas the overall effect of response group on means of the response variable were based on omega squared. For Cohen's D, the following classification was used: 0 to <0.2 = no effect; 0.2 to <0.5 = small effect (*); 0.5 to <0.8 = medium effect (**); >0.8 = large effect (***).

[3]Group means are displayed for continuous variables.

Item Level Missing Data:
Almost all health plans (96%) had less than 5% missing response to the osteoporosis item among survey responders (See Table 7). The average missing item rate across health plans was 2% (See Table 8).

Table 7: Frequency of Missing Data for Osteoporosis Item (HEDIS 2012 data)

Missing %	Frequency	Percent	Cumulative Freq	Cumulative %
< 5%	477	96.36	477	96.36
5% - <10%	12	2.42	489	98.79
10% - <15%	5	1.01	494	99.8
>= 15%	1	0.2	495	100

Table 8: Distribution of Non-response of Osteoporosis Item Across Plans (HEDIS 2012 data)

Description	N	Mean	std	p10	p25	p50	p75	p90
Missing	495	2.17%	1.71%	0.58%	1.21%	1.90%	2.69%	3.65%

N: number of plans
std: standard deviation
p: percentile

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*)

2014 submission:

Survey Level Missing Data Analysis: In general, late responders tended to be similar to nonresponders on every variable except institutionalization. Responders and non-responders tended to be similar in terms of age, health (HCC score), and gender. While most differences between groups were small, there were moderate differences seen between responders and non-responders with regard to the percent of individuals who were non-white, had dual eligibility, or were institutionalized. Analysis of the effect size showed none of these differences to be large or strong. It is not surprising that individuals with dual eligibility, disability or in institutions are less likely to respond. This population likely has a higher rate of cognitive impairment. The dual eligible population is also more likely to be non-English speaking (the mailed survey is offered in English, Spanish, and Chinese). Overall, our measurement advisory panel did not feel these differences reflected significant non-response survey bias.

Item Level Missing Data Analysis: The overall frequency of missing data for the OTO questions was very low across plans. Over 96% of plans had 5% or fewer missing responses for the OTO question. Only one plan was missing data for the OTO question for more than 15% of their survey population. The distribution showed that on average across plans, survey responses were missing data on the OTO question 2.2% of the time. Based on this analysis, it is unlikely that missing data on this question would bias performance results.

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.**

Other

If other: Information is gathered through the Health Outcome Survey (HOS).

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**.

No data elements are in defined fields in electronic sources

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).

The Health Outcomes Survey is conducted through mailed surveys with telephone follow-up. There are currently no plans to conduct this survey over the web or in an electronic form. There is concern that somemany Medicare beneficiaries do not have access to a computer or internet to complete the survey in electronic format. There is also a concern that moving to an internet-based mode of administration will bias results, as older frail adults may be less likely to complete the survey using an internet mode. Given the nature of the questions in the HOS survey, there is also a high priority to ensure confidentiality of the results.

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

Attachment:

3c. Data Collection Strategy

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

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

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

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA requires all Medicare plans to contract with an NCQA-certified HOS survey vendor to administer the survey. NCQA developed its Survey Vendor Certification Program to establish standardization of data collection and thereby promote comparability of results across Medicare health plans. NCQA provides oversight for Health Outcome Survey implementation and prohibits survey vendors from augmenting or adjusting the HOS protocol or instrument, expect as approved by NCQA and CMS. Oversight includes the following elements:

1. Quality Assurance Plan from the survey vendor focused on protocol adherence and implementation of corrective actions and evaluation of their impact on performance

2. Bi-weekly reporting from survey vendors about the data-collection process
3. Site visits for selected survey vendors
4. Offsite monitoring or survey vendor correspondence with respondents, telephone interviews, data record review and other elements

In addition to the HEDIS Survey Vendor Certification, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.

Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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 Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Health Plan Ratings http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Medicare Advantage Reporting http://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Health Plan Ratings http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Medicare Advantage Reporting http://www.cms.gov/Medicare/Prescription-Drug-

	Coverage/PrescriptionDrugCovGenIn/PerformanceData.html Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Health Plan Ratings http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017

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

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

NCQA STATE OF HEALTH CARE QUALITY ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported by WedMD and on the NCQA website. These ratings are based on a plan''s performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.

MEDICARE ADVANTAGE DISPLAY PAGE: This measure is listed on the display page for Medicare Advantage (Medicare Part C). This means that while performance on this measure is not tied to incentives; plans have the option to report.

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?*)
N/A

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.*)
N/A

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. The HOS survey is administered to members of health plans with at least 500 beneficiaries. The survey utilizes random sampling for health plans with more than 1,200 members. Additional population descriptions and sampling methods are described in Section S.15. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure as it is successfully collected through the Medicare Health Outcome Survey.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs.

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.**
During the measure's last major update in 2014, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure.

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.
Current HEDIS rates indicate that just under three quarters of women over the age of 65 in Medicare Advantage plans report having received at least one bone mineral density test in their lifetime. In 2015, the spread in national health plan performance was 60.0 to 86.5 percent (10th to 90th percentiles).

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.**
There were no identified unexpected findings for this measure during testing or since implementation.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits for this measure during testing or since implementation.

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)**
0046 : Screening for Osteoporosis for Women 65-85 Years of Age
0053 : Osteoporosis Management in Women Who Had a Fracture
2417 : Risk Assessment/Treatment After Fracture

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.**

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.)
There are multiple NQF-endorsed measures of osteoporosis prevention and management. During the last measure update in 2014, we undertook a comprehensive harmonization exercise to align several NQF-endorsed osteoporosis measures where possible given the different measure focus, methods of data collection and level of accountability. Below we describe the harmonization between this measure (0037) and the most closely related measure, 0046.

Measure 0046 assesses the percentage of women who have a bone mineral density test to screen for osteoporosis. Measure 0046 is collected using medical record review and is only specified for physician level reporting). The rationale for different data sources is the availability of data for the level of reporting.

Measure 0037 is a health plan level measure. Since the recommended timeframe for osteoporosis testing is at least once since turning age 65 or prior to age 65 if at risk, the measure is specified as “ever” having a bone mineral density test. It is not feasible for a Medicare Advantage plan to have access to enough historical claims data or medical record data to determine if its entire member population has ever had a bone mineral density test. Therefore, a survey method is the recommended data source for collecting this type of historical data.

Measure 0046 is a physician level measure. Physicians are limited by the same lack of historical data, but also have limited resources to field and collect a survey of their patient population. Therefore, this measure looks for documentation in the medical record that a bone mineral density test was performed. This documentation may come from previous medical records requested by the current physician on past care.

The harmonized measure elements described below are reflective of the most recent measure versions submitted for endorsement.

Harmonized Measure Elements between 0037 and 0046:
- Type of Test: Because measure 0037 is a survey measure, the term “bone mineral density test” is used to refer to dual energy x-ray

absorptiometry test. The simplified term is used because cognitive testing indicated it was more understandable to survey respondents. We have harmonized the two measures by ensuring both measures only capture testing done of the hip or spine; however 0046 is able to capture more specificity about the type of test done due to the data source used for measure collection.

- Eligible Population: Both measures are focused on women age 65-85 years of age.

- Timeframe for testing: Both measures address whether testing was done at least once in the woman's lifetime.

Given the two different data sources, we do not expect the two measures (0037 and 0046) to have exactly comparable results; however, the two measures address the same quality gap for different levels of accountability.
- Measure 0037 addresses whether a health plan is addressing the risk for osteoporosis in the patient population by determining the percent of the population that had a bone mineral density test regardless who their provider is. This test may have been done outside of the context of their primary care provider.
- Measure 0046 addresses whether individual providers are addressing the risk for osteoporosis in their patient population by determining if an individual had a bone mineral density test to screen for osteoporosis and if their provider is aware of those results and can advise on appropriate risk reduction.

Measure 0053 addresses a different population than 0037 (i.e., women who have experienced a fragility fracture), and is therefore focused on secondary prevention of future fractures as opposed to screening for osteoporosis. Measure 2417 also focuses on those who had a fragility fracture and then received secondary prevention. Therefore, we consider these measures to be related but not competing. The differences between these measures are reflective of the different guidelines for general population screening and secondary prevention. Where it is appropriate to the measure focus and evidence, we have aligned the 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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
Geriatric Measurement Advisory Panel (GMAP)
Wade Aubry, MD, University of California, San Francisco
Arlene S Bierman, MD, MS, Agency for Healthcare Research and Quality (AHRQ)
Patricia A. Bomba, MD, MACP, Excellus BlueCross BlueShield
Nicole Brandt, PharmD, MBA, BCGP, BCPP, FASCP, University of Maryland, School of Pharmacy
Jennie Chin Hansen, RN, Geriatric Expert
Joyce Dubow, MUP, Consumer Representative
Gustavo Ferrer, MD, Aventura Hospital
Peter Hollmann, MD, University Medicine
Jeffrey Kelman, MD, MMSc, Centers for Medicare & Medicaid Services (CMS)
Karen Nichols, MD, AmeriHealth Caritas Family of Companies Steven Phillips, MD, CMD, Geriatric Specialty Care

Jane Sung, JD, AARP
Eric G Tangalos, MD, FACP, AGSF, CMD, Mayo Clinic
Dirk Wales, MD, PsyD, Cigna HealthSpring
Joan Weiss, PhD, RN, CRNP, Health Resources and Services Administration
Neil Wenger, MD, UCLA Division of General Internal Medicine and RAND

Osteoporosis Advisory Workgroup
Joyce Dubow, MUP, Consumer Representative
Margery Gass, MD, NCMP, The North American Menopause Society
Peter Hollmann, MD, University Medicine
Steven Petak, MD, MACE, JD, Endocrinologist, Houston Methodist Hospital Academic Associates
Kenneth G. Saag, MD, MSc, Divison of Clinical Immunology and Rheumatology, University of Alabama at Birmingham

Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
Jonathan Darer, MD, MPH, Medicalis
Helen Darling, MA, City of Washington, DC
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare & Medicaid Services
David Grossman, MD, MPH, Kaiser Permanente Washington
Christine S. Hunter, MD, US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Nancy Lane, PhD, Newton, MA
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric Schneider, MD, MSc, FACP, The Commonwealth Fund
Marcus Thygeson, MD, MPH, San Rafael, CA
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms
Lina Walker, PhD, AARP

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2003

Ad.3 **Month and Year of most recent revision:** 05, 2014

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if clinical guidelines or evidence has changed significantly

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

www.ncqa.org/publications.

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0046
Measure Title: Screening for Osteoporosis for Women 65-85 Years of Age
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: Percentage of women 65-85 years of age who ever had a central dual-energy x-ray absorptiometry (DXA) test to check for osteoporosis.
Developer Rationale: This measure assesses the number of women 65-85 who have ever received a dual-energy x-ray absorptiometry (DXA) test to check for osteoporosis. There is convincing evidence that bone mineral density tests predict short-term risk for osteoporotic fractures. There is also evidence osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

Numerator Statement: The number of women who have documentation in their medical record of having received a DXA test of the hip or spine.
Denominator Statement: Women age 65-85.
Denominator Exclusions:
Diagnosis of osteoporosis at the time of the encounter.
Patient receiving hospice services anytime during the measurement period.

Measure Type: Process
Data Source: Electronic Health Data, Electronic Health Records, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: May 01, 2007 **Most Recent Endorsement Date:** Dec 30, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- The developer briefly described the link between bone mineral density and the patient's health outcomes in reduced risk of developing osteoporosis or sustaining a fragility fracture and reduced risk of morbidity and mortality.
- The developer provided a draft US Preventive Services Task Force Recommendation (release April 9,2018) including recommendations for the following:
  - "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool."
  - Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
  - The developer summarized the Quality, Quantity, and Consistency of the body of evidence associated with the draft US Preventive Services Task Force Recommendation (2018).

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided an updated (although still Draft) US Preventive Services Task Force Recommendation (released April 9, 2018) which continues to support their measure focus.

Exception to evidence

NA

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*
  - *If derived from patient report, does the target population value the measured process or structure and find it meaningful?*

Guidance from the Evidence Algorithm

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

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

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

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

Performance Data:

- Developer provided performance data from Physician Quality Reporting System (PQRS) data from 2009 -2012,. The mean performance rate in 2012 was 58.7%. In 2012, 505,070 eligible providers (6.1%) chose to report on this measure.
- The mean performance rate in 2009 was 56.1%; 2010 was 55.1%; and 2011 was 61.2%.

Disparities:

- Developer did not provide disparities from the measure. However cited a national cohort study by Gillespie and Morin that examined claims data from 2008 to 2014 for trends in osteoporosis screening in women age 50 and older. The data was categorized based on race/ethnicity, age, sex, and socioeconomic status.
    - They found that after controlling for other factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001).
    - After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening.
    - Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis.
- In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015).

Questions for the Committee:

- *Specific questions on information provided for gap in care.*
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The evidence has been updated and is stronger than when the measure was previously reviewed. There is no need to repeat the discussion and to vote on the evidence again.
- I agree that since the evidence is updated and directional, we do not need to have a discussion and vote on the Evidence
- Data is extracted from PQRS 2012 reports
- 2017 Review (Viswanathan et al 2017) Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. I am not aware of any other studies.
- I am not aware of any evidence or studies related to this care area other than those referenced by the Developer (as follows):
    - The developer provided a draft US Preventive Services Task Force Recommendation (release April 9,2018).
- There is strong evidence that DXA testing reliably diagnoses osteoporosis and assesses fracture risk and that pharmacologic rx reduces fracture risk.
- No studies address potential harms of DXA testing. The risk from using pharmacologic agents is small (USPTF)
- This is a process measure of women ages 65-85 without the diagnosis of osteoporosis who have evidence of ever having a DEXA scan. The additional evidence is the draft USPSTF recommendation (4/9/18, Grade B) for DEXA screening at age 65 and over or earlier if risk factors are present. The relationship between DEXA scores in the osteoporotic range and the risk of significant fractures is moderate.

- Process measure - Yes/no to population having a DXA scan or evidence of DXA scan during study period. Directly relates to measure.
- The evidence provided supports screening – is once in a lifetime enough? Would it never need to be repeated between the ages of 65 – 85?
- New information from USPSTF (draft April 2018) which is updated, directionally same, adds to body of evidence which supports the rationale (BMT predicts short term risk for fx; tx decreases fx for women identified as risk; fx decreased health, function and possibly death)
- Relationship to pt outcomes: see rationale
- Strength of evidence: moderate
- Evidence applicable to process of care being measured: yes
- Therefore no need for repeat discussion and vote on evidence 1a.

Performance Gap

- There is a performance gap, based on the measure data provided. It is interesting to note however that there has been no significant change over the years reported.
- I agree that since the evidence is updated and directional, we do not need to have a discussion and vote on the Evidence.
- Data is extracted from PQRS 2012 reports.
- Yes. Less than optimal performance (55.1-61.2%). No disparities data on the measure provided. Disparities: Yes. But two studies showed non-hispanic black women (18.2%) least likely to have screening compared with other categories (22-22.7%). Lower socioeconomic status- lower rates of screening.
- The Developer cited performance gaps including opportunities for improvement as follows:
    - Developer provided performance data from Physician Quality Reporting System (PQRS) data from 2009 - 2012,. The mean performance rate in 2012 was 58.7%. In 2012, 505,070 eligible providers (6.1%) chose to report on this measure.
    - The mean performance rate in 2009 was 56.1%; 2010 was 55.1%; and 2011 was 61.2%.
- The Developer did not independently review disparities. However, the Developer did cite:
    - Developer did not provide disparities from the measure. However cited a national cohort study by Gillespie and Morin that examined claims data from 2008 to 2014 for trends in osteoporosis screening in women age 50 and older. The data was categorized based on race/ethnicity, age, sex, and socioeconomic status.
    - They found that after controlling for other factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001).
    - After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening.
    - Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis.
    - In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015).
- Performance Data is not current, i.e., not cited since 2012
- Performance Data is also limited: 6.1% of eligible providers(30,000) reported on this measure in 2012. Also the interquartile difference between the 25th and 75th percentile is 77.3% which demonstrates considerable variability.
- Performance data since the measure's submission in 2014 was not provided. Medical literature published since 2014 has some studies indicating disparities in DEXA scanning by race and socioeconomic status exists.
- Performance gap in 2012 was 77% for IQR (25-75). Average performance is ~58% in 2012. Data from 6% of providers eligible to use this measure in PQRS reporting.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- Performance data: 2009 56.1%, 2010 55.1%, 2011 61.2%, 2012 58.7% so room for improvement
- Racial and socioeconomic disparities evidence from research (not done by developer) suggest need for measure

- Rating: moderate

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

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: Primary Care and Chronic Illness project team staff

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

Questions for the Committee regarding reliability:
- o *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- o *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:
- o *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- o *The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

Committee pre-evaluation comments

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

Reliability Specification
- Inclusions, Exclusions (hospice and those already with a dx of osteoporosis) reported are clear and there a re no concerns regarding consistency in measurement.

- No concerns about specifications
- Code/value set is clear.
- No concerns.
- The measure submission in 2014 used PQRS data from 2009-2012. Updated PQRS data was not included in this submission and PQRS reporting ended in 2016, evolving to MIPS.
- Measures clearly defined, algorithm clear. Kappa = 0.77 (substantial reliability)
- Concur with the analysis of the staff evaluator.
- No concerns
- Specifications are precise, unambiguous, complete so can be consistently implemented
- Empirical reliability testing was done at pt level data elements and captured all critical data elements (from 2014 data Kappa on scoring was .77)
- Reliabiilty rating: moderate

Reliability Testing

- No concerns related to reliability.
- None, if the QPP and PQRS are reported similarly.
- There was high inter-rater reliability 90% (numerator) to 100% (denominator) which demonstrates that the data can be accurately extracted from charts. (kappa score of .77)
- No concerns
- No concerns with the reliability of the measure.
- No
- There is no information about how the numbers of women ever screened by DEXA scans can be determined from MIPS data or any other substitute source of such clinical data.
- Initial reliability testing was done with individual chart reviews by reviewers. Current methods of data submission include digital data from EHR, which is not addressed in the reliability testing.
- Concur with the analysis of the staff evaluator.
- no concerns

Validity Testing

- There are no updates to any of the components of validity since the 2014 report.
- No concerns.
- No concerns.
- 2014-Large gap in performance between 25th and 75th percentiles.
- 2014: No missing data
- No concerns with missing data. Is set up to be primarily administrative in nature including the use of electronic health records.
- There is no information about how the numbers of women ever screened by DEXA scans can be determined from MIPS data or any other substitute source of such clinical data.
- Face validity - good
- Concur with the analysis of the staff evaluator.
- Face validity testing only (not empirical validity) from 2014 and no updates
- Only available data is from CMS Quality Payment Program (formerly PQRS)
- Followed PCPI process for measuring face and content validity and concluded that measure meets PCPI standards for actually measuring what it purports to measure and allows users to make conclusions about quality of care that is provided
- Demonstrated face validity (did not have empirical validity testing)
- Face validity results demonstrate sufficient agreement, and this is maintenance measure
- Therefore, can rate this as moderate for validity

Other Threats to Validity

- No concerns
- Exclusions are logical and consistent
- This measure is not risk adjusted
- Risk adjustment grayed out: Why?
- Exclusions: Seems appropriate. May need a quick discussion
- No concerns with measure exclusions.
- No threats to validity from Exclusions or lack of Risk Adjustment.
- Excluded groups are appropriate. Does not account for severe mobility issues causing challenges for patients to

get to facility and get onto DXA table.
- Concur with the analysis of the staff evaluator.
- NA

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

Data Specifications and Elements

- The measure has Information is gathered through multiple data sources (administrative data, electronic clinical data, and paper records).
- Some data elements are in defined fields in electronic sources.
- Developer shared no difficulties on the use of this measure.
- This is not an eMeasure

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

Feasibility

- Highly feasible as data is easily extracted from administrative claims, PQRS/QPP and electronic health records sources. The measure developer also encourages its use without cost.
- Data elements are generated by chart abstraction etc.
- Only some data elements are in defined fields in electronic sources.
- No significant concerns. However, no clear path to eCQM noted. Data collected from administrative data, electronic data and paper abstraction. However, "feedback on the use of this measure has been positive, with few questions raised by participating clinicians…"
- Data is collected using electronic health records and claims. No concerns with feasibility of data collection.
- The required data elements are routinely generated and used during care delivery
- There is no information about how the numbers of women ever screened by DEXA scans can be determined from MIPS data or any other substitute source of such clinical data.
- Places responsibility on provider to find previous test results if DXA done in a previous practice or ordered by another provider outside the practice. Does not account for patient refusals for DXA.
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- According to NCQA, data sources are administrative, electronic and paper clinical to allow for widespread reporting across clinical practices (ie. Some physicians may not use electronic records, and paper record review is more cumbersome but still feasible)
- Few questions raised by providers and issues reviewed biweekly by developer with CMS
- Apparently not all data elements are in defined fields within EHR;

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
QUALITY PAYMENT PROGRAM: this measure is used in the quality payment program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

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
This measure uses the following methods to obtain input: including vetting of the measure with measure advisory panels including NCQA's Geriatric Measurement Advisory Panel and NCQA's Osteoporosis Advisory Workgroup during a re-evaluation process in 2014.

The measure is in the Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS) by CMS. CMS solicits feedback and has a designated space on their webpage with information on how to share feedback with them. The measure owner has not received any feedback on this measure.

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

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

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

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

4b. Usability 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 developer states that performance rates increased by 2.6 % from 2009-2012, which show minor improvement. The measure is not required, however developer hope rate will show improvement as there is increased accountability to report this measure.

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

- Per developer, there is a possibility that the measure may result in overuse of DXA testing for women.
    - The measure looks for documentation that a DXA test was performed. If a provider does not have access to previous medical records documenting that a DXA was performed or patient reported/provided results of a previous DXA, then a repeat DXA may be ordered even if the patient had a previous DXA.
    - There is no guidance on how frequently a woman should receive a test, but the USPTSF recommends that a minimum two-year gap is needed to detect bone density changes between tests.
    - This measure also has the potential to lead women who had a bone mineral density test prior to 65 to repeat screening after age 65, which may not be indicated by the woman's risk factors.

Potential harms [potential harms]

- See unexpected findings section above.
- In addition, per the evidence form submitted by developer, developer noted the following by USPSTF: "The USPSTF found no studies that described harms of screening for osteoporosis in men or women. Based on the nature of screening with bone measurement tests and the low likelihood of serious harms, the USPSTF found adequate evidence to bound these harms as no greater than small. Harms associated with screening may include radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)."

Additional Feedback:

- In 2015 NQF Endocrine Report, the Committee mentioned the following concerns about the usability of the measure:
    - concern about no time limitation on the measure, meaning that any bone mineral density test done over the course of a women's lifetime would meet the requirements of the measure.
    - concern over the difficulty in obtaining medical records for patients who had the study performed in the more distant past, particularly when under the care of another provider.
    - concern that overuse of the bone mineral density testing may be an unintended consequence of the measure

***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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- This measure is currently used in public reporting and accountability programs.
- Public and payment reporting through CMS, Quality payment program, formerly PQRS
- It is selected as a reporting measure by 6.1 % of eligible providers.
- CMS QPP-public reporting.

- No new feedback-only 80 EPs reported on this measure in 2015!
- QUALITY PAYMENT PROGRAM: this measure is used in the quality payment program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).
- This measure uses the following methods to obtain input: including vetting of the measure with measure advisory panels including NCQA's Geriatric Measurement Advisory Panel and NCQA's Osteoporosis Advisory Workgroup during a re-evaluation process in 2014.
- The measure is in the Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS) by CMS. CMS solicits feedback and has a designated space on their webpage with information on how to share feedback with them. The measure owner has not received any feedback on this measure.
- Performance data is publicly reported.
- Clinicians being measured are provided feedback from NCQA
- There is no information about how the numbers of women ever screened by DEXA scans can be determined from MIPS data or any other substitute source of such clinical data. However, Screening for Osteoporosis for Women 65-85 years of age is one of the quality measures under MIPS that can be selected for reporting.
- Feedback is provided by payers who ask for data reporting, or who are measuring this as part of P4V programs.
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity.
- CMS QPP
- 2015 looked at scores from 80 EPs
- deemed a priority measure by NCQA geriatric and osteoporosis advisory workgroups
- not a required measure for CMS QPP (6.1% of EPs reported on this measure)

Usability

- There are no concerns with respect to known unintended consequences.
- Among physicians that reported on the measure to CMS, between 2009 and 2012 screening increased only 2.6 %.
- No. "Currently, this measure is not required for physician reporting (they have the option). There is hope that with increasing accountability to report on this measure then the rate will begin to show improvement.". No update on "progress or improvement in usability" reported.
- Some unintended overutilization theoretically possible, but no data available.
- No identified harm in the use of this measure. Outcomes could be improved if the measure is reported.
- QUALITY PAYMENT PROGRAM: this measure is used in the quality payment program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).
- Public reporting of performance results should lead medical professionals to seek better percentages in future years.
- As screening with DEXA scans in this population is suggested by the USPSTF at a Grade B level, securing and following the numbers of scans if accurately determined would be considered a benefit to the selected population and a benefit to individuals in the population. There was concern in the original submission for the possibility of promoting overuse of such scans. A complicating factor is the relatively recent decrease in health insurance coverage for DEXA scans, which appears to be significant in reducing the women in this age group who receive screening.
- Benefits far outweigh harms. Additional testing may be done for providers to meet the measure to avoid finding results from past tests.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- no concenrns
- Potential overuse if provider does not have prior records or if testing prior to age 65

Criterion 5: Related and Competing Measures

Related or competing measures

- Developer identified six related or competing measures
    - 0037 : Osteoporosis Testing in Older Women (OTO)

- 0045 : Communication with the physician or other clinician managing on-going care post fracture for men and women aged 50 years and older
- 0048 : Osteoporosis: Management Following Fracture of Hip, Spine or Distal Radius for Men and Women Aged 50 Years and Older
- 0053 : Osteoporosis Management in Women Who Had a Fracture
- 2416 : Laboratory Investigation for Secondary Causes of Fracture
- 2417 : Risk Assessment/Treatment After Fracture

Harmonization

- Related/Competing Measures:
  - Measure 0046 and 0037 have the same measure focus and same target population. However, they both have different levels of analysis and accountability, and use different data sources. Measure 0046 is collected using medical record review and is only specified for physician level reporting, whereas 0037 is collected using the Health Outcome Survey to patients for a health plan level measure. The rationale for different data sources is the availability of data for the level of reporting. The developer describes in further detail the harmonization of these two measures in/since 2014. Both measures have same steward (NCQA). (Competing)
  - Measure 0045, 0048, 0053, 2416, and 2417 address a different population than 0046.
    - These measures address women who have experienced a fracture, and are focused on secondary prevention of future fractures as opposed to screening for osteoporosis. (Related)
    - Per developer, where it is appropriate to the measure focus and evidence they have aligned the measures.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0046

Measure Title: Screening for Osteoporosis for Women 65-85 Years of Age

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

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

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

   ☒Yes (go to Question #2)

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

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

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

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

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☐Yes (go to Question #4)

☒No (skip Questions #4-5 and go to Question #6)

4. 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.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☐Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***
   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:
    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)
    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.
***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?
***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

No updated data from 2014 testing
75th percentile scoring at 100%

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☐Yes (go to Question #18)

    ☒No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☐Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☐Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☐Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with <u>patient-level data elements</u>?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Prior validity studies of the same data elements may be submitted*

    ☐Yes (go to Question #22)

    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*

REFERENCE: *Testing attachment, section 2b1.*
TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☒Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.
TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

REFERENCE: Testing attachment, section 2c
***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

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

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0046

Measure Title: Screening for Osteoporosis for Women 65-85 Years of Age

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

Date of Submission: 4/9/2018

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: Screening for Osteoporosis

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

2014 Submission

Female patients at risk for osteoporosis (age 65 and older)>>> bone mineral density test to check for low bone mass or osteoporosis >>> low bone mass identified >>> patient evaluated for treatment options >>> treatment >>> reduced risk of developing osteoporosis or sustaining a fragility fracture >>> maintained quality of life and reduced risk of morbidity and mortality.

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

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

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

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

USPSTF Recommendation: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission NCQA acknowledges that as of April 9, 2018, the U.S. Preventive Services Task Force (USPSTF) has released a DRAFT recommendation statement for osteoporosis screening. A draft Evidence Review was also published in November 2017. When published, NCQA will evaluate the final recommendation statement and supporting evidence review and consider any potential changes that may be needed for this measure. However, based on the draft recommendation statement we do not anticipate that any major revisions will be needed. U.S. Preventive Services Task Force. 2017. *Draft Recommendation Statement: Osteoporosis to Prevent Fractures: Screening*. https://www.uspreventiveservicestaskforce.org/Page/Document/draft-recommendation-statement/osteoporosis-screening1 U.S. Preventive Services Task Force. 2017. *Draft Evidence Review: Osteoporosis to Prevent Fractures: Screening.* https://www.uspreventiveservicestaskforce.org/Page/Document/draft-evidence-review/osteoporosis-screening1 2014 Submission U.S. Preventive Services Task Force. 2011. Screening for osteoporosis: US preventive services task force recommendation statement. Annals of internal medicine, 154(5), 356. http://www.uspreventiveservicestaskforce.org/uspstf10/osteoporosis/osteors.htm, accessed May 2, 2014.
Quote the guideline or	2018 Submission "The USPSTF recommends screening for osteoporosis with bone measurement testing to

recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool." 2014 Submission "The USPSTF recommends screening for osteoporosis in women aged 65 years or older and in younger women whose fracture risk is equal to or greater than that of a 65-year-old white woman who has no additional risk factors."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission The USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in women age 65 years and older is at least moderate.
Provide all other grades and definitions from the evidence grading system	2018 Submission N/A
Grade assigned to the **recommendation** with definition of the grade	2018 Submission Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. 2014 Submission This measure is based on a grade B recommendation from the USPSTF. Grade B: The USPSTF recommends the services. 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	2018 Submission Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. 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. Grade I: 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. 2014 Submission

	Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. 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?	2018 Submission The DRAFT evidence report (Viswanathan et al 2017) supporting this guideline outlines the quantity and quality of evidence, which are summarized below for the key questions of the 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? • As in the previous 2011 review, found no good or fair quality studies eligible for this key question Key Question 2a. What is the accuracy and reliability of screening approaches to identify adults who are at increased risk for osteoporotic fracture? • Accuracy of Clinical Risk Assessment Tools for Identifying Osteoporosis: included 37 articles (35 studies, fair or good quality) • Accuracy of Bone Measurement Tests Used to Identify Low Bone Mass and Osteoporosis: included 11 studies, fair or good quality • Accuracy of Bone Measurement Tests Used to Predict Fracture: included 21 studies, fair or good quality • Accuracy of Fracture Risk Prediction Instruments: included 1 systematic review and 13 fair or good quality observational studies Key Question 2b. What is the evidence to determine screening intervals and how do these vary by baseline fracture risk? • Included 2 articles (2 studies, good quality) Key Question 3. What are the harms of screening for osteoporotic fracture risk? • Found no eligible studies that addressed this question Key Question 4a. What is the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality? Bisphosphonates: • Alendronate: included 7 studies, fair or good quality • Zoledronic Acid: included 2 studies, fair or good quality • Risedronate: included 4 studies, fair or good quality • Etidronate: included 2 fair quality studies • Ibandronate: identified no studies or trials that assessed the benefits of ibandronate for preventing fractures

Raloxifene:

- Included 1 large good quality RCT

Estrogen:

- No studies included

Denosumab:

- Included 3 fair quality trials

Parathyroid Hormone:

- Included 2 fair quality trials

Key Question 4b. How does the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality vary by subgroup, specifically in postmenopausal women, premenopausal women, men, younger age groups (age <65 years), older age groups (age ≥65 years), baseline bone mineral density, and baseline fracture risk?

Bisphosphonates:

- Zoledronic Acid, Etidronate, Ibandronate: found no relevant results in included studies for subgroup analysis
- Alendronate: included 1 study
- Risedronate: included 1 RCT

Raloxifene:

- Included 1 study

Estrogen:

- No studies included

Denosumab:

- Included 1 fair quality trial

Parathyroid Hormone:

- Included 1 fair quality trial

Key Question 5. What are the harms associated with pharmacotherapy?

Bisphosphonates:

- Alendronate: included 16 studies, fair or good quality
- Zoledronic Acid: included 4 studies, fair or good quality
- Risedronate: included 4 studies, fair or good quality
- Etidronate: included 2 fair quality studies
- Ibandronate: included 7 fair quality studies

Raloxifene:

- Included 6 studies

Estrogen:

- No studies included

Denosumab:

- Included 3 fair quality studies

Parathyroid Hormone:

- Included 2 fair quality studies

Viswanathan, M., et al. 2017. "Screening to Prevent Osteoporotic Fractures: An Evidence Review for the U.S. Preventive Services Task Force." Available here: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryDraft/osteoporosis-screening1

2014 Submission

	Quantity: N/A (not required for previous submission) Quality: N/A (not required for previous submission)
Estimates of benefit and consistency across studies	2018 Submission The following text is quoted directly from the USPSTF recommendation statement. The USPSTF found no studies that evaluated the effect of screening for osteoporosis on fracture rates or fracture-related morbidity or mortality. The USPSTF found convincing evidence that bone measurement tests are accurate for detecting osteoporosis and predicting osteoporotic fractures in women and men. The USPSTF found adequate evidence that clinical risk assessment tools are moderately accurate in identifying risk of osteoporosis and osteoporotic fractures. The USPSTF found convincing evidence that drug therapies reduce subsequent fracture rates in postmenopausal women. The benefit of treating screening-detected osteoporosis is at least moderate in women age 65 years and older and younger postmenopausal women who have similar fracture risk. The harms of treatment range from no greater than small for bisphosphonates and parathyroid hormone to small to moderate for raloxifene and estrogen. Therefore, the USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in these groups of women is at least moderate. The USPSTF concludes that the evidence is inadequate to assess the effectiveness of drug therapies in reducing subsequent fracture rates in men without previous fractures. Treatments that have been proven effective in women cannot necessarily be presumed to have similar effectiveness in men, and the direct evidence is too limited to draw definitive conclusions. Thus, the USPSTF could not assess the balance of benefits and harms of screening for osteoporosis in men. **2014 Submission** N/A
What harms were identified?	2018 Submission The following is quoted directly from the USPSTF draft recommendation statement: "The USPSTF found no studies that described harms of screening for osteoporosis in men or women. Based on the nature of screening with bone measurement tests and the low likelihood of serious harms, the USPSTF found adequate evidence to bound these harms as no greater than small. Harms associated with screening may include radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)." **2014 Submission** N/A
Identify any new studies conducted since the SR. Do	2018 Submission To our knowledge, there have been no published studies since the systematic review that

the new studies change the conclusions from the SR?	would impact the recommendations. **2014 Submission** N/A

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0046
Corresponding Measures:
De.2. **Measure Title:** Screening for Osteoporosis for Women 65-85 Years of Age
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** Percentage of women 65-85 years of age who ever had a central dual-energy x-ray absorptiometry (DXA) test to check for osteoporosis.
1b.1. **Developer Rationale:** This measure assesses the number of women 65-85 who have ever received a dual-energy x-ray absorptiometry (DXA) test to check for osteoporosis. There is convincing evidence that bone mineral density tests predict short-term risk for osteoporotic fractures. There is also evidence osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

S.4. **Numerator Statement:** The number of women who have documentation in their medical record of having received a DXA test of the hip or spine.
S.6. **Denominator Statement:** Women age 65-85.
S.8. **Denominator Exclusions:** Diagnosis of osteoporosis at the time of the encounter.
Patient receiving hospice services anytime during the measurement period.

De.1. **Measure Type:** Process
S.17. **Data Source:** Electronic Health Data, Electronic Health Records, Paper Medical Records
S.20. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: May 01, 2007 **Most Recent Endorsement Date:** Dec 30, 2014

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

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0046_-_Evidence.docx
1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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 assesses the number of women 65-85 who have ever received a dual-energy x-ray absorptiometry (DXA) test to check for osteoporosis. There is convincing evidence that bone mineral density tests predict short-term risk for osteoporotic fractures. There is also evidence osteoporosis treatment reduces the incidence of fracture in women who are identified to be at risk of an osteoporotic fracture. Fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
The following data were extracted from Physician Quality Reporting System (PQRS) and reflect claims data for services provided from January 1, 2012 through December 31, 2012 . PQRS is a pay-for-reporting incentive program that allows providers to choose which quality measures to report on. The program has been renamed as the Quality Payment Program. In 2012, of 505,070 eligible providers, 6.1% chose to report on this measure. Performance data is summarized at the physician level and described by mean, 10th, 25th, 50th, 75th and 90th percentile.

This measure has been updated since these data were collected. Therefore, these data reflect performance on the previous version of the measure which looked for either screening or treatment for osteoporosis.

Performance Rate for all Reporting Providers for 2012
Mean | 10th | 25th | 50th | 75th | 90th
58.7%|0.00% | 22.7% | 64.3% | 100% | 100%

The following data (also extracted from PQRS) show the average performance rates for several years prior to 2012.

Average performance rates from 2009-2011
2009 | 56.1%
2010 | 55.1%
2011 | 61.2%

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.
N/A

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.*
Performance data stratified by different variables is not currently available for this measure based on how it is reported in the CMS Quality Payment Program (QPP). However, if demographic variables were collected accurately this measure could be stratified by things such as race/ethnicity or other factors, in order to assess the presence of health care disparities.

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
There is a misconception that osteoporosis is only a concern for non-Hispanic white women, which may result in delaying prevention and treatment in non-White and Hispanic populations. African-American and Hispanic women are less likely to believe they are at risk for osteoporosis (NIH NIAMS 2010). In a national cohort study (Gillespie and Morin 2017), researchers examined medical claims data from 2008 to 2014 for trends in osteoporosis screening in women age 50 and older. They found that after controlling for other

factors, non-Hispanic Black women were least likely to have osteoporosis screening (18.2%) compared with other racial/ethnic categories (range: 22.0%-22.7%, P<.001). After controlling for various patient characteristics, non-Hispanic Asian and Hispanic women in the 50-64 and 65-79-year age groups had the highest odds of screening. Outside of racial and ethnic disparities, women with lower socioeconomic status had lower rates of screening for osteoporosis (Gillespie and Morin). In a retrospective cohort study, researchers from the University of California, Davis Health Systems also found that Black women and women with more socioeconomic barriers were less likely to be screened for osteoporosis (Amarnath et al 2015). Interventions that target population screening are needed to improve the rates of osteoporosis screening for women age 65 and older.

Amarnath, A. L. D., Franks, P., Robbins, J. A., Xing, G., & Fenton, J. J. (2015). Underuse and overuse of osteoporosis screening in a regional health system: a retrospective cohort study. Journal of general internal medicine, 30(12), 1733-1740.
Gillespie, C. W., & Morin, P. E. (2017). Trends and disparities in osteoporosis screening among women in the United States, 2008-2014. The American journal of medicine, 130(3), 306-316.
National Institutes of Health. (2010). National Institute of Arthritis and Musculoskeletal and Skin Disorders. Osteoporosis and African American Women. Accessed at: www.niams.nih.gov/hi/topics/osteoporosis/opbkgr.htm

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):*
Endocrine, Musculoskeletal, Musculoskeletal : Osteoporosis

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

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

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.)*
NA

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

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

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Since the last measure update, we have incorporated an exclusion for patients in hospice. It would not be beneficial to assess older women in hospice care to see whether they had a bone mineral density test to screen for osteoporosis.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The number of women who have documentation in their medical record of having received a DXA test of the hip or spine.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Documentation of a central dual-energy x-ray absorptiometry (DXA) test ever being performed.

The numerator criteria is met by documentation in the medical record that the patient has had a central dual-energy x-ray absorptiometry test. This measure is also collected in the Quality Payment Program using the following codes specific to the quality measure:
Performance Met: G8399 Patient with documented results of a central Dual-energy X-Ray Absorptiometry (DXA) ever being performed.
Performance Not Met: G8400 Patient with central Dual-energy X-Ray Absorptiometry (DXA) results not documented, reason not given.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women age 65-85.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Women who had a documented patient encounter (see Table 1 for encounter codes) during the reporting period.

Table 1: Patient encounter during the reporting period (CPT): 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Diagnosis of osteoporosis at the time of the encounter.
Patient receiving hospice services anytime during the measurement period.

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.)*
The denominator exclusion criteria is met by documentation in the medical record of a diagnosis of osteoporosis at the time of the encounter (see Table 2 for diagnosis codes).

Table 2: Diagnosis of osteoporosis on date of encounter (ICD-10-CM): M80.00XA, M80.00XD, M80.00XG, M80.00XK, M80.00XP, M80.00XS, M80.011A, M80.011D, M80.011G, M80.011K, M80.011P, M80.011S, M80.012A, M80.012D, M80.012G, M80.012K, M80.012P, M80.012S, M80.019A, M80.019D, M80.019G, M80.019K, M80.019P, M80.019S, M80.021A, M80.021D, M80.021G,

M80.021K, M80.021P, M80.021S, M80.022A, M80.022D, M80.022G, M80.022K, M80.022P, M80.022S, M80.029A, M80.029D, M80.029G, M80.029K, M80.029P, M80.029S, M80.031A, M80.031D, M80.031G, M80.031K, M80.031P, M80.031S, M80.032A, M80.032D, M80.032G, M80.032K, M80.032P, M80.032S, M80.039A, M80.039D, M80.039G, M80.039K, M80.039P, M80.039S, M80.041A, M80.041D, M80.041G, M80.041K, M80.041P, M80.041S, M80.042A, M80.042D, M80.042G, M80.042K, M80.042P, M80.042S, M80.049A, M80.049D, M80.049G, M80.049K, M80.049P, M80.049S, M80.051A, M80.051D, M80.051G, M80.051K, M80.051P, M80.051S, M80.052A, M80.052D, M80.052G, M80.052K, M80.052P, M80.052S, M80.059A, M80.059D, M80.059G, M80.059K, M80.059P, M80.059S, M80.061A, M80.061D, M80.061G, M80.061K, M80.061P, M80.061S, M80.062A, M80.062D, M80.062G, M80.062K, M80.062P, M80.062S, M80.069A, M80.069D, M80.069G, M80.069K, M80.069P, M80.069S, M80.071A, M80.071D, M80.071G, M80.071K, M80.071P, M80.071S,M80.072A, M80.072D, M80.072G, M80.072K, M80.072P, M80.072S, M80.079A, M80.079D, M80.079G, M80.079K, M80.079P, M80.079S, M80.08XA, M80.08XD, M80.08XG, M80.08XK, M80.08XP, M80.08XS, M80.80XA, M80.80XD, M80.80XG, M80.80XK, M80.80XP, M80.80XS, M80.811A, M80.811D, M80.811G, M80.811K, M80.811P, M80.811S, M80.812A, M80.812D, M80.812G, M80.812K, M80.812P, M80.812S, M80.819A, M80.819D, M80.819G, M80.819K, M80.819P, M80.819S, M80.821A, M80.821D, M80.821G, M80.821K, M80.821P, M80.821S, M80.822A, M80.822D, M80.822G, M80.822K, M80.822P, M80.822S, M80.829A, M80.829D, M80.829G, M80.829K, M80.829P, M80.829S, M80.831A, M80.831D, M80.831G, M80.831K, M80.831P, M80.831S, M80.832A, M80.832D, M80.832G, M80.832K, M80.832P, M80.832S, M80.839A, M80.839D, M80.839G, M80.839K, M80.839P, M80.839S, M80.841A, M80.841D, M80.841G, M80.841K, M80.841P, M80.841S, M80.842A, M80.842D, M80.842G, M80.842K, M80.842P, M80.842S, M80.849A, M80.849D, M80.849G, M80.849K, M80.849P, M80.849S, M80.851A, M80.851D, M80.851G, M80.851K, M80.851P, M80.851S, M80.852A, M80.852D, M80.852G, M80.852K, M80.852P, M80.852S, M80.859A, M80.859D, M80.859G, M80.859K, M80.859P, M80.859S, M80.861A, M80.861D, M80.861G, M80.861K, M80.861P, M80.861S, M80.862A, M80.862D, M80.862G, M80.862K, M80.862P, M80.862S, M80.869A, M80.869D, M80.869G, M80.869K, M80.869P, M80.869S, M80.871A, M80.871D, M80.871G, M80.871K, M80.871P, M80.871S, M80.872A, M80.872D, M80.872G, M80.872K, M80.872P, M80.872S, M80.879A, M80.879D, M80.879G, M80.879K, M80.879P, M80.879S, M80.88XA, M80.88XD, M80.88XG, M80.88XK, M80.88XP, M80.88XS, M81.0, M81.6, M81.8

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

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 = Higher 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.)*
Step 1: Determine the eligible population. To do so, identify patients who meet all the specified criteria.
-Sex: Females
-Age: 65-85 years of age
-Patient encounter during the reporting period (12 months)

Step 2: Exclude from the eligible population in step 1 patients who have a diagnosis of osteoporosis at time of encounter.

Step 3: Identify the number of patients with a central dual-energy x-ray absorptiometry test documented.

Step 4: Calculate the rate (number of patients who had a central dual-energy x-ray absorptiometry test documented divided by the eligible population).

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

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

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

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 Data, Electronic Health Records, Paper Medical 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.
This measure is based on administrative claims to identify the eligible population and medical record documentation collected in the course of providing care to health plan patients to identify the numerator. In the Quality Payment Program this measure is coded using G-codes specific to quality measurement.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, 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.)*
N/A

2. Validity – See attached Measure Testing Submission Form
0046_-_Testing_Form_v7.1-636588800587376811.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.
No

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.
No

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.
No - This measure is not risk-adjusted

Measure Number (*if previously endorsed*)**:** 0046

Measure Title: Screening for Osteoporosis for Women 65-85 Years of Age

Date of Submission: 4/9/2018

Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record

☐ claims	☐ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ 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*).

1.3. What are the dates of the data used in testing?
2014 Submission:
Sample 1: Testing of data element reliability was performed during field testing in 2009.
Sample 2: Testing of performance variability was performed using 2012 performance data from the Physician Quality Reporting System.

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*)
2014 Submission:
Sample 1: This measure was tested for data element reliability using field test data. To identify clinics for field testing, the American Academy of Orthopedic Surgeons (AAOS) posted an announcement online and also identified practices that were known through their previous work with the AAOS. Of the thirteen clinics who expressed an interest in the field-testing, two were chosen to participate. These two sites were chosen based on having participated in the 2009 Physician Quality Reporting Initiative (PQRI) program with additional consideration given to balancing practice size, location, and use of an EHR or paper medical record. One site was located in New Mexico and one was located in South Carolina.

Sample 2: This measure is used in the Physician Quality Reporting System (PQRS) as a performance measure for eligible professionals. 2012 performance data from PQRS was used to examine the variation in performance for this measure. The number of providers submitting data for this measure in 2012 was 35,079.

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*)
2014 Submission:

Sample 1: Desired sample size for testing was calculated for this measure with 0.80 power, 0.05 significance, and testing for a kappa of substantial agreement (0.8) versus moderate agreement (0.4). Expected performance was conservatively assumed at 0.5. Based on these assumptions and calculations, the minimum number of patients needed for the sample was 28. Both sites included 30 patients in their samples.

Sample 2: The number of patients eligible to be reported on in PQRS for 2012 was 13,339,356.

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.
2014 Submission:
Sample 1 was used to test data element reliability. Sample 2 was used to demonstrate performance variation.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
2018 Submission:
We did not analyze performance by social risk factors.

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*)
2014 Submission:
Critical data element reliability: Reliability was tested by assessing whether two abstractors, reviewing the same full medical (including both inpatient and outpatient notes), would come to the same conclusion as to the patient meeting the measure, not meeting the measure, or qualifying as an exception. Two abstractors independently assessed whether patients met numerator inclusion criteria for each case that met denominator inclusion criteria. Following the data abstraction, the mismatches were tallied. Agreement between abstractors was measured using the kappa statistic (a measure of agreement adjusted for agreement that can occur by chance).

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*)
2014 Submission:
Denominator: Agreement between the two independent reviewers was 100% for the denominator data element. The Kappa was undefined for this data element as both abstractors agreed that all of the cases met denominator eligibility criteria.

Exceptions: Agreement between the two independent reviewers was 100% for the exception data element. The Kappa was undefined for this data element as both abstractors agreed that none of the cases should be excluded.

Numerator: Agreement between the two reviewers was 90% with agreement that the numerator criteria was met in 19/30 cases and not met in 8/30 cases. The reviewers disagreed about 3/30 cases where one reviewer found evidence that the numerator criteria was met and one review did not find evidence in the medical record that numerator criteria was met. The abstractors then reconciled the mismatches through an adjudication process and determined 22/30 cases met numerator criteria. A Kappa statistic was calculated to demonstrate the degree of agreement adjusted for chance (K=0.77; 95% CI: 0.63-1.00).

Table 1 below displays the overall agreement for the all the measure components combined. Concordance between the abstractors is 90% with moderate agreement above what would be expected (K=0.77; 95% CI 0.53-1.00).

Table 1: Inter-rater reliability of measure components

		Reviewer B				
		Not Study Eligible	Not Met	Met	Exception	Total
Reviewer A	Not Study Eligible	0	0	0	0	0
	Not Met	0	8	0	0	8
	Met	0	3	19	0	22
	Exception	0	0	0	0	0
	Total	0	11	19	0	30

"Not Study Eligible" means that the denominator criteria were not met.
"Not Met" means denominator criteria were met, numerator criteria were not met and exceptions (exclusions) did not apply.
"Met" means denominator criteria were met and numerator criteria were met.
"Exception" means denominator criteria were met, numerator criteria were not met and exceptions applied.

Kappa Coefficient	0.77
Kappa LL (95% Confidence Interval)	0.53
Kappa UL (95% Confidence Interval)	1.00
Observed Agreement Rate	0.90
Expected Agreement Rate	0.56

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?*)
2014 Submission:
Interpretation of data element reliability testing: The below scale was used in the field test to interpret the kappa score. The numerator had a kappa score of .77, which indicates that there was substantial agreement that the two abstractors came to the same conclusion as to patients who met the numerator. This suggests the measure elements can be reliably abstracted from medical records.

Kappa	Strength of Agreement
0.00	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate

0.61 – 0.80 Substantial
0.81 – 0.99 Almost perfect

Landis, J.R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159—174

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

2018 Submission:

There are no updates to the validity testing for this measure since the last submission. The only publicly available data for this measure are from reporting in the CMS Quality Payment Program, however these data are not constructed in a way that allows NCQA to test empirical validity of the measure.

2014 Submission:

Critical Data Element Validity: The testing conducted for this measure by the AMA/Physician Consortium for Performance Improvement (PCPI) is described above under "Reliability." This testing demonstrates inter-rater reliability of two reviewers using the same measure specification to draw conclusions from the same "gold-standard" data source (e.g. medical record). Reliability testing demonstrated that two independent reviewers looking at the same full medical record had high agreement on every data element and the overall performance measure score. We believe this testing demonstrates not only data element reliability but also validity, that is to say the accuracy of the measure specification to identify all data elements from the medical record.

Assessment of face validity: This measure was also evaluated for face-validity by the AMA/PCPI which oversees the measure development process of clinically relevant physician-level performance measures. To assess the face validity of measures, PCPI follows a standardized process for measure development which includes:

- Convening cross-specialty, multidisciplinary work groups to assess the face and content validity of each measure. The groups establish the measure's ability to capture what it is designed to capture using a consensus process that consists of input from multiple stakeholders, including practicing physicians and experts with technical measure expertise.
- Review of the evidence, gaps in care and potential for impact of the measure:
  - Consider existing guideline recommendations and the strength of evidence
  - Consider gaps in care, variation, cost and frequency data
- Posting the draft measure for a 30-day public comment period. The PCPI solicits feedback from PCPI members, quality improvement collaboratives, providers, consumers, public/private purchasers and others with an interest in the measure.
- The PCPI work group reviews comments received, revises and modifies the draft performance measures as deemed appropriate by the work group. The public comments and responses are posted to the PCPI website as part of the voting process.

- Final vote by PCPI members eligible to vote. The PCPI encourages all voting member organizations to vote so the required quorum is met.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
2014 Submission:
Results of face validity assessment: This measure was reviewed and developed by a joint work group that included experts in osteoporosis treatment as well as representatives from the following organizations: American Academy of Family Physicians; American Academy of Orthopaedic Surgeons; American Association of Clinical Endocrinologists; American College of Rheumatology; The Endocrine Society; American Medical Association; National Osteoporosis Foundation; National Committee for Quality Assurance; and The Joint Commission. The joint work group members came to consensus on the final recommended specification for this measure in October 2006. See section Ad. 1. Workgroup/Expert Panel Involved in Measure Development for a list of participants of the Osteoporosis Work Group.

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?*)
2014 Submission:
Interpretation of face validity assessment: These results indicate that the multiple experts and stakeholders concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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*)

2018 Submission:
The exclusion for this measure is based on clearly specified codes that indicate the patient received hospice services during the measurement period. While this code has not been specifically tested in the context of this measure, it is considered valid for identifying patients who receive hospice services. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons.

2014 Submission:
N/A – no exclusions

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*)
2018 Submission:
NA

2014 Submission:
N/A – no exclusions

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)
2018 Submission:
NA

2014 Submission:
N/A – no exclusions

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

2b3.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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

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

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

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:
☐ **Published literature**
☐ **Internal data analysis**
☐ **Other (please describe)**

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

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

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

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

2014 Submission:

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

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*)

2014 Submission:

2012 Variation in Performance across Providers

Mean Rate	EP	10th	25th	50th	75th	90th	IQR
58.7%	326,372	0.00%	22.7%	64.3%	100.0%	100.0%	77.3

EP: Number of patients meeting denominator criteria across all providers submitting data to the Physician Quality Reporting System on this measure

IQR: Interquartile range

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?*)

2014 Submission:

The results above indicate there is a large gap in performance between providers at the 25th and 75th percentiles. This demonstrates a large variation in performance and significant room for improvement on this measure for many providers. It should be noted that performance data from the PQRS program does not reflect performance system wide because physicians have the option to report. We look forward to more detailed performance reports from PQRS that may demonstrate longitudinal provider-specific performance improvements.

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
2014 Submission:
This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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*)
2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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*)

2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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**.

Some data elements are in defined fields in electronic sources

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).

To allow for widespread reporting across physicians and clinical practices, this measure in practice is collected through multiple data sources (administrative data, electronic clinical data, and paper records).

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

Attachment:

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the**

measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Feedback on use of this measure in CMS PQRS program has been positive with few questions raised by participating clinicians to the CMS vendor. NCQA works with the CMS vendor to review any questions or issues raised with the measure on a bi-weekly basis.

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)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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 CMS Quality Payment Program https://qpp.cms.gov CMS Quality Payment Program https://qpp.cms.gov Payment Program CMS Quality Payment Program https://qpp.cms.gov CMS Quality Payment Program https://qpp.cms.gov

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

QUALITY PAYMENT PROGRAM: this measure is used in the quality payment program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

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?*)

N/A

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.*)
N/A

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.
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): In 2015, 80 eligible professionals (EP) reported on the measure. EPs submitting PQRS data to CMS received a PQRS feedback report on whether they satisfactorily reported and if they are subject to a payment adjustment. The data in these reports may help EPs determine whether or not it is necessary to submit an informal review request. An informal review is a process that allows EPs to request a review of their payment adjustment determination.

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.**
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): Each year, QPP individual EPs and QPP group practices receive feedback reports on whether they satisfactorily reported and if they are subject to the future downward payment adjustment. CMS hosts training sessions on these reports and posts audio recording and slide presentations on their webpages. CMS also provides technical assistance and maintains webpages with information about accessing and understanding these reports.

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.
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): CMS solicits feedback and has a designated space on their webpage with information on how to share feedback with them. The measure owner has not received any feedback on this measure.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): No feedback was received specific to this measure.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure went through a re-evaluation process in 2014. During that process, feedback on the measure was obtained from measure advisory panels including NCQA's Geriatric Measurement Advisory Panel and NCQA's Osteoporosis Advisory Workgroup. This measure was deemed a priority measure by the panels.

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.**
At the time of the measure's last major update in 2014, no feedback had been received.

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.

From 2009-2012 the average performance rate increased by 2.6 percent, which shows minor improvement amongst those providers who choose to report on this measure. In 2012, of 505,070 eligible providers, 6.1% chose to report on this measure.

Currently, this measure is not required for physician reporting (they have the option). There is hope that with increasing accountability to report on this measure then the rate will begin to show improvement.

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.**

There is a possibility that the measures may result in overuse dual-energy x-ray absorptiometry (DXA) testing for women. The measure looks for documentation that a DXA test was performed. If a provider does not have access to previous medical records documenting that a DXA was performed or patient reported/provided results of a previous DXA, then a repeat DXA may be ordered even if the patient had a previous DXA. There is no guidance on how frequently a woman should receive a test, but the USPTSF recommends that a minimum two-year gap is needed to detect bone density changes between tests. This measure also has the potential to lead women who had a bone mineral density test prior to 65 to repeat screening after age 65, which may not be indicated by the woman's risk factors.

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

N/A

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)**

0037 : Osteoporosis Testing in Older Women (OTO)
0045 : Communication with the physician or other clinician managing on-going care post fracture for men and women aged 50 years and older
0048 : Osteoporosis: Management Following Fracture of Hip, Spine or Distal Radius for Men and Women Aged 50 Years and Older
0053 : Osteoporosis Management in Women Who Had a Fracture
2416 : Laboratory Investigation for Secondary Causes of Fracture
2417 : Risk Assessment/Treatment After Fracture

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.**

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.)
Although 0037 and 0046 have the same measure focus and same target population they are specified for different levels of analysis and accountability, and use different data sources. We have described above where the measures are conceptually harmonized and the rationale for where the measures cannot be harmonized in their technical specifications due to the level of analysis and data source.

RESPONSE TO 5a.2 (insufficient space above):
There are multiple NQF-endorsed measures of osteoporosis prevention and management. In the most recent update, we undertook a comprehensive harmonization exercise to align several NQF-endorsed osteoporosis measures where possible given the different measure focus, methods of data collection and level of accountability. Below we describe the harmonization between this measure (0046) and the most closely related measure, 0037.

Measure 0046 assesses the percentage of women who have a bone mineral density test to screen for osteoporosis. Measure 0046 is collected using medical record review and is only specified for physician level reporting. The rationale for different data sources is the availability of data for the level of reporting.
- Measure 0037 is a health plan level measure. Since the recommended timeframe for osteoporosis testing is at least once since turning age 65 or prior to age 65 if at risk, the measure is specified as "ever" having a bone mineral density test. It is not feasible for a Medicare Advantage plan to have access to enough historical claims data or medical record data to determine if the entire member population ever had a bone mineral density test. Therefore a survey method is the recommended data source for collecting this type of historical data.
- Measure 0046 is a physician level measure. Physicians are limited by the same lack of historical data, but also have limited resources to field and collect a survey of their patient population. Therefore, this measure looks for documentation in the medical record that a bone mineral density test was performed. This documentation may come from previous medical records requested by the current physician on past care.
The harmonized measure elements described below are reflective of the most recent measure versions submitted for endorsement.

Harmonized Measure Elements between 0037 and 0046:
- Type of Test: Because measure 0037 is a survey measure, the term "bone mineral density test" is used to refer to "dual energy x-ray absorptiometry test." This term is used because cognitive testing indicated the term was more understandable to survey respondents. We have harmonized the two measures by ensuring both measures only capture testing done of the hip or spine; however, 0046 is able to capture more specific about the type of test done due to the data source used for measure collection.
- Eligible Population: Both measures are focused on women age 65-85 years of age.
- Timeframe for testing: Both measures address whether testing was done at least once in the woman's lifetime.

Given the two different data sources, we do not expect the two measures (0037 and 0046) to have exactly comparable results; however, the two measures address the same quality gap for different levels of accountability.
- Measure 0037 addresses whether a health plan is addressing the risk for osteoporosis in the patient population by determining the percent of the population that had a bone mineral density test regardless who their provider is. This test may have been done outside of the context of their primary care provider.
- Measure 0046 addresses whether individual providers are addressing the risk for osteoporosis in their patient population by determining if an individual had a bone mineral density test to screen for osteoporosis and if their provider is aware of those results and can advise on appropriate risk reduction.

Measures 0045, 0048, 0053, 2416, and 2417 address a different population than 0046. These measures address women who have

experienced a fracture, and are focused on secondary prevention of future fractures as opposed to screening for osteoporosis. Therefore, we consider these measures to be related but not competing. The differences between these measures are reflective of the different guidelines for general population screening and secondary prevention. Where it is appropriate to the measure focus and evidence we have aligned the 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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
Geriatric Measurement Advisory Panel (GMAP)
Wade Aubry, MD, University of California, San Francisco
Arlene S Bierman, MD, MS, Agency for Healthcare Research and Quality (AHRQ)
Patricia A. Bomba, MD, MACP, Excellus BlueCross BlueShield
Nicole Brandt, PharmD, MBA, BCGP, BCPP, FASCP, University of Maryland, School of Pharmacy
Jennie Chin Hansen, RN, Geriatric Expert
Joyce Dubow, MUP, Consumer Representative
Gustavo Ferrer, MD, Aventura Hospital
Peter Hollmann, MD, University Medicine
Jeffrey Kelman, MD, MMSc, Centers for Medicare & Medicaid Services (CMS)
Karen Nichols, MD, AmeriHealth Caritas Family of Companies Steven Phillips, MD, CMD, Geriatric Specialty Care
Jane Sung, JD, AARP
Eric G Tangalos, MD, FACP, AGSF, CMD, Mayo Clinic
Dirk Wales, MD, PsyD, Cigna HealthSpring
Joan Weiss, PhD, RN, CRNP, Health Resources and Services Administration
Neil Wenger, MD, UCLA Division of General Internal Medicine and RAND

Osteoporosis Advisory Workgroup
Joyce Dubow, MUP, Consumer Representative
Margery Gass, MD, NCMP, The North American Menopause Society
Peter Hollmann, MD, University Medicine
Steven Petak, MD, MACE, JD, Endocrinologist, Houston Methodist Hospital Academic Associates
Kenneth G. Saag, MD, MSc, Divison of Clinical Immunology and Rheumatology, University of Alabama at Birmingham

Physician Consortium for Performance Improvement Osteoporosis Workgroup
Steven Petak, MD, MACE, JD, Endocrinologist, Houston Methodist Hospital Academic Associates
Kenneth G. Saag, MD, MSc, Divison of Clinical Immunology and Rheumatology, University of Alabama at Birmingham
Robert Alder, MD
H. Chris Alexander, III, MD, FACP
Donald Bachman, MD, FACR
Joel Brill, MD
Jan Busby-Whitehead, MD

Thomas Dent, MD
Nancy Dolan, MD
Leonie Gordon, MB, ChB
Thomas Griebling, MD
Richard Hellman, MD, FACP, FACE
Marc C. Hochberg, MD, MPH
C. Conrad Johnston, Jr. MD
Joseph Lane, MD
Leon Lenchik, MD
Bonnie McCafferty, MD, MSPH
Michael Maricic, MD
Michael L. O'Dell, MD, MSHA, FAAFO
Sam J.W. Romeo, MD, MBA
Frank Salvi, MD, MS
Joseph Shaker, MD
Madhavi Vemireddy, MD
David Wong, MD, MSc, FRS(C)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2003

Ad.3 **Month and Year of most recent revision:** 05, 2014

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if clinical guidelines or evidence has changed significantly

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.6 **Copyright statement:** The performance measures and specifications were developed by and are owned by the National Committee for Quality Assurance ("NCQA"). The performance measures and specifications are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these materials and can rescind or alter these materials at any time. These materials may not be modified by anyone other than NCQA. Anyone desiring to use or reproduce the materials without modification for an internal, quality improvement non-commercial purpose may do so without obtaining any approval from NCQA. All other uses, including a commercial use and/or external reproduction, distribution and publication must be approved by NCQA and are subject to a license at the discretion of NCQA.

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

Content reproduced with permission from HEDIS, Volume 2: Technical Specifications for Health Plans. To purchase copies of this publication, including the full measures and specifications, contact NCQA Customer Support at 888-275-7585 or visit www.ncqa.org/publications.

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0053
Corresponding Measures:
Measure Title: Osteoporosis Management in Women Who Had a Fracture
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of women age 50-85 who suffered a fracture and who either had a bone mineral density test or received a prescription for a drug to treat osteoporosis.
Developer Rationale: The intent of this measure is secondary prevention of fractures through the appropriate diagnosis and treatment of osteoporosis. Detecting osteoporosis and initiating treatment will help to prevent future fractures from occurring. Future fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

Numerator Statement: Patients who received either a bone mineral density test or a prescription for a drug to treat osteoporosis after a fracture occurs.
Denominator Statement: Women who experienced a fracture, except fractures of the finger, toe, face or skull. Three denominator age strata are reported for this measure:
Women age 50-64
Women age 65-85
Women age 50-85
Denominator Exclusions: - Exclude women who had a bone mineral density test during the 24 months prior to the index fracture.
- Exclude women who had a claim/encounter for osteoporosis treatment during 12 months prior to the index fracture.
- Exclude women who received a dispensed prescription or had an active prescription to treat osteoporosis during the 12 months prior to the index fracture.
- Exclude women who are enrolled in a Medicare Institutional Special Needs Plan (I-SNP) or living long-term in an institution any time during the measurement year.
- Exclude women receiving hospice care during the measurement year.**:**

Measure Type: Process
Data Source: Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Dec 30, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 or **Summary of prior review in [year]**

- The developer provided updated U.S. Preventative Task Force (USPSTF) guideline (2017 Draft) recommendation statement for osteoporosis screening.
    - "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool."
    - The guideline was assigned **Grade B.**
        - The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
    - The developer summarizes the Quality, Quantity, and Consistency of the body of evidence associated with the guidelines.
- The developer provided updated guidelines from the American Association of Clinical Endocrinologists (AACE) (2016).
    - "Who needs to be screened for osteoporosis? All postmenopausal women > 50 years should undergo clinical assessment for osteoporosis and fracture risk, including a detailed history and physical examination." (page 7)
    - "The AACE recommends bone mineral density testing for women aged 65 and older and younger postmenopausal women at increased risk for bone loss and fracture based on fracture risk analysis." (page 10)
    - The guideline was assigned **Grade B.**
        - Evidence from at least 1 large well-designed clinical trial, cohort or case-controlled analytic study, or meta-analysis.
        - No conclusive level 1 publication; ≥ 1 conclusive level 2 publications demonstrating benefit > risk. (see Table 2 in section above)
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guidelines, as this information was not available in the guidelines.
        - To supplement the guidelines, the developer cited three systematic reviews. These citations were provided in the 2014 submission of the measure.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided updated U.S. Preventative Task Force (USPSTF) (2017 Draft) and American Association of Clinical Endocrinologists (AACE) (2016) guidelines. The grade of the both updated guidelines did not differ from the previous version.

Exception to evidence

NA

Questions for the Committee:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*
  - *If derived from patient report, does the target population value the measured process or structure and find it meaningful?*

Guidance from the Evidence Algorithm
Process measure with systematic review (Box 3) ->Summary of the QQC provided (Box 4) ->Systematic review concludes moderate quality evidence.

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

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

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

Health plan level:

- Developer provided performance data extracted from HEDIS data collection for Medicate Advantage Health Plans from 2014, 2015, and 2016.
  - Mean: 35.9% (2014) to 40.0% (2016)
  - Standard Deviation: 17.3% (2014) to 19.0% (2016)

Clinician level:

- Developer provided performance data extracted from Physician Quality Reporting System (PQRS), from 2009-2011.
  - Mean: 56.5% (2009) to 70.6% (2011)
- In 2012, of 204,369 eligible providers, only 0.8% chose to report on this measure. Therefore, the performance rates below are reflective of less than one percent of Medicare providers. At the time of data collection this measure applied to women age 50 and older.
  - Mean: 70.0% (2012)
- In 2014 the measure was revised to reflect the added upper age limit. For the next year of quality measurement reporting, the physician level performance will be reported for the 50-85 age strata.
  - The developer did not provide performance data.

Disparities

- NCQA does not currently collect performance data stratified by race, ethnicity, or language. CMS does not currently report data currently reported by data stratified by different variables in the PQRS, now QPP program.
- The developer cited a study that found differences the prevalence of osteoporosis among people of all ethnic backgrounds (Silverman, 1988).
- The developer cited two studies that suggests that African American women receive less dual-energy x-ray absorptiometry screenings and treatment for osteoporosis. (Hamrick, 2016, 2012)
- The developer cited a cohort study that found African American women had the lowest treatment rates for osteoporosis when compared with women of other races. (Liu, 2016).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The evidence provided includes updated evidence and is stronger than when this measure was previously reviewed. There is no need for the committed to repeat the discussion and to vote on the evidence.
- I agree that since the evidence is updated and directional, we do not need to have a discussion and vote on the Evidence
- Data is extracted from PQRS 2012 reports
- I am not aware of any changes in the evidence base or new studies/information other than what was submitted by the Developer.
- The developer provided updated U.S. Preventative Task Force (USPSTF) guideline (2017 Draft) recommendation statement for osteoporosis screening.
- The developer provided updated guidelines from the American Association of Clinical Endocrinologists (AACE) (2016).
- The developer has provided updated evidence which strengthens this measure. Women who fracture are at higher risk of additional fractures. A bone density test guides the provider to treat with pharmaceutical agents providing secondary prevention. Avoiding additional fractures improves quality of life and survival. The evidence is Moderate
- USPSTF recommendation cites potential harm associated with patient time and effort – later referred to as "opportunity costs" – what are the opportunity costs and what is our threshold for acceptance related to these costs alone and in concert with other competing medical costs. Seems this should be quantified and measured to determine the harm isn't too great.
- This is a process measure of the percentage of women age 50-85 who suffered a fracture and who either had a bone mineral density test or received a prescription for a drug to treat osteoporosis within 6 months of the fracture using HEDIS data. This measure continues in the HEDIS 2018 measures set (women 65 years old or older living long term in an institutional setting are excluded). NQF Measure 0053 additionally excludes women who had a bone density test within 24 months of the index fracture, those who have had an encounter or prescription for osteoporosis within 12 months of the index fracture, and those in hospice care. HEDIS data through 2016 is included. The developer added the draft USPSTF recommendation (4/9/18, Grade B) for DEXA screening in women at age 65 and older (or earlier if risk factors are present) and the AACE guideline (2016, Grade B) that all postmenopausal women 50 years old and older should have a clinical assessment for osteoporosis and fracture risk and women age 65 and older (or earlier if risk factors are present) should have bone mineral density testing. The AACE guideline strongly recommends pharmacologic therapy for patients with osteopenia or low bone mass and a history of fragility fracture of the hip or spine. The 2018 submission includes reference to a literature review, Viswanathan, M., et al. 2017. "Screening to Prevent Osteoporotic Fractures: An Evidence Review for the U.S. Preventive Services Task Force." Published studies addressing the question of the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality are listed by therapy agent. The USPSTF found convincing evidence that drug therapies reduce subsequent fracture rates in postmenopausal women. The USPSTF draft recommendation for DEXA screening states that the benefit of treating screening detected osteoporosis is at least moderate in women age 65 years and older and younger postmenopausal women who have similar fracture risk. The harms of treatment range from no greater than small to moderate. This does not specifically address treatments initiated after an index fracture.
- Women from 50 to 85 who suffered fracture and had subsequent BMT and/or treatment with osteoporosis medication
- Rationale is secondary prevention of future fractures as these cause decreased health, decreased function and possibly mortality
- Developer provided updated information from USPSTF and AAC Endocrinologists which is directionally same and enhances support
- Relationship to outcomes: see rationale
- Strength of evidence: moderate
- Evidence applicable to the process of care being measured? Yes
- Therefore no need for repeat discussion and vote on evidence

- There is moderate evidence to suggest that for fragility fractures predict the risk of future fractures. Treatment with Osteoporosis medications can reduce the risk of future fractures.

Performance Gap

- Based on the evidence provided, there is indeed a performance gap. With this in mind, there is a need for this national performance measure.
- There is a large gap with the data demonstrating less than 605 performance as per electronic chart, billing data and paper chart data.
- Disparity Data is not available through PQRS/QPP. Following provided from publications evidence.
    - Black women and women of low SES and SES barriers had lower rates of screening.
    - Asian and Hispanic women had the highest screening rates when controlling for various external factors.
- Developer provided performance data extracted from HEDIS data collection for Medicate Advantage Health Plans from 2014, 2015, and 2016. The results showed an opportunity for health plan level improvement.
- The Developer provided performance data extracted from Physician Quality Reporting System (PQRS), from 2009-2011. The results showed an opportunity for physician level improvements.
- The Developer cited two studies that suggested disparity in identification or care/treatment for Women of African American descent.
- For Medicare Advantage Plans performance data extracted from HEDIS data collection 2014, 2015, and 2016 showed a Mean: 35.9% (2014) to 40.0% (2016); Standard Deviation: 17.3% (2014) to 19.0% (2016). Improvement is needed
- Clinician level: performance data extracted from PQRS, from 2009- 2011. Mean: 56.5% (2009) to 70.6% (2012)
- In 2012 only 0.8% chose to report on this measure, reflective of <1% of Medicare providers. No current data for Clinicians presented by developer. There are likely disparities in care for this measure but this was not addressed by developer.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- Performance data from Medicare Advantage Health Plans are included through 2016. Clinician data from PQRS for years 2009-2012 are included. Both sets indicate significant performance gaps. There is no data from years later than 2016. Performance was not analyzed by social risk factors.
- Racial disparities identified in levels of BMT
- Rating: moderate
- Mean performance at the health plan level is no better than 40%. From PQRS data, physician level performance mean was 70%, however, less than 1% of eligible providers opted to report on this measure.
- Data suggest that African-American women are tested less frequently for osteoporosis and treated less frequently as compared to non-African-American women. This data, however, does not come from this NCQA collected data.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability (Health Plan Level): Specifications and Testing
2a. Reliability (Clinician Level): Specifications and Testing
2b. Validity (Health Plan Level): Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability, Missing Data
2b. Validity (Clinician Level): Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability, Missing Data

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is

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

Validity

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

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

Composite measures only:

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

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

Evaluation of Reliability and Validity (and composite construction, if applicable): Staff evaluation

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.)?*
- *The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Guidance from the Reliability Algorithm
Health Plan Level:
Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing (Box 4) →Appropriate method (Box 5) → High certainty that measure results are reliable (Box 6a) → High

Clinician Level:
Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing not conducted (Box 4) → Data elements tested (Box 8) → Appropriate method (Box 9) → Moderate level of agreement between raters (Box 10a) → Moderate

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

Guidance from the Validity Algorithm
Health Plan Level:
Threats to Validity (Box 1) → Empirical validity testing with measure as specified (Box 2) → Score-level testing (Box 5) →Appropriate method (Box 6) → Moderate certainty or confidence that measure results are valid (Box 7b) → Moderate

Clinician Level:
No empirical validity testing conducted. Developer provided justification.

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

Committee pre-evaluation comments

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

Reliability Specification

- Inclusions, Exclusions (hospice and those already with a dx of osteoporosis) reported are clear and there are no concerns regarding consistency in measurement.
- Codes and value sets are clearly defined. No concerns with the reliability of the specifications.
- Specifications are clearly defined despite multiple data sources.
- Concur with the analysis of the staff evaluator.
- Reliability testing of HEDIS measures was described in the original application. There is no additional material provided in 2018.
- No concerns
- Therefore no need for discussion and vote
- No concerns

Reliability Testing

- I have no concerns related to the reliability of this measure
- None, if the QPP and PQRS are reported similarly and health plan data is included.
- There was high inter-rater reliability 90% (numerator) to 100% (denominator) which demonstrates that the data can be accurately extracted from charts. (kappa score of .77)
- No concerns about the reliability testing of the measure.
- No
- Concur with the analysis of the staff evaluator.
- No
- No concerns
- Therefore no need for discussion and vote
- No concerns

Validity Testing

- I have no concerns on the threats to validity of this measure.
- There are no updates to any of the components of validity since the 2014 report.
- No concerns with the validity or validity testing.
- No concerns with Validity Testing.
- Concur with the analysis of the staff evaluator.
- The 2018 submission provides updated material on the process by which NCQA evaluates face validity.
- No concerns
- Used interquartile range for performance with 10.9% gap in performance between 1st and 3rd quartiles
- No concerns.

Other Threats to Validity

- Exclusions are logical and consistent
- This measure is not risk adjusted
- No identified threats to validity according to the information provided by the Developer.
- No threats to validity related to exclusions or risk adjustment
- Concur with the analysis of the staff evaluator.
- No concerns; they included empirical validity testing
- Admin data and clinical data correlated
- 2018 submitted face validity testing
- NCQA audit process ensures calculations not biased due to missing data
- No concerns; exclusions based upon clearly specified codes which are considered valid for identifying patients who should be excluded;
- 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 measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records). While only some data elements are in defined fields in electronic sources, the elements are generated as byproduct of care processes. This measure is also a HEDIS measure and NCQA conducts audits to verify that HEDIS specifications are met.
- This is not an eMeasure.

Questions for the Committee:

- o *Are the required data elements routinely generated and used during care delivery?*
- o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- o *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- The required data elements are routinely generated and used during care delivery and reimbursement.
- Highly feasible as data is easily extracted from administrative claims, PQRS/QPP and electronic health records sources. The measure developer also encourages its use without cost.
- The measure is feasible. It uses both claim/encounter data and chart data. I tis more human resource intensive and more expensive to collect data. Use of electronic health records for performance measure reporting purposes may have a positive impact on resource use and cost.
- Data elements are routinely generated and used during care delivery.
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- The initial submission describes the HEDIS Audit process. It mentions that physician feedback has been positive for PQRS but that program changed to MIPS after 2016.
- NCQA utilizes HEDIS audit to ensure that measure users are following NCQA standards in reporting for health plans
- At clinician level, reporting is done through CMS QPP and NCQA works with CMS vendors to ensure no issues are raised (biweekly)
- Capturing data within a EHR may be challenging if Tests are performed in another health system. Use of claims data can overcome this challenge.

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

HEALTH PLAN LEVEL USE:

- STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
- CMS MEDICARE ADVANTAGE STAR RATING: This measure is included in the composite Medicare Advantage Star Rating.
- HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
- NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.
- NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans.
- NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance.

PHYSICIAN LEVEL USE

- CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

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

- Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.

Additional Feedback: NA

***Questions for the Committee**:*

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

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

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

4b. Usability 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

- Health Plan Level: From 2014 to 2016, the average performance rate has increased by four percentage points. Since 2013, rates have increased about 18.4 percent for health plans in the 90th percentile (see section 1b.2 for summary of data from health plans). In 2016, a total of 277 Medicare health plans reported data on this measure. These data are nationally representative.
- Physician Level: From 2009-2012 the average performance rate has increased by 13.5 percent, which shows steady improvement amongst those providers who chose to report on this measure. In 2012, there were 204, 369 eligible providers who were able to report on this measure and only 0.8% choose to report. Therefore, the 2012 average performance rate is reflective of less than one percent of Medicare providers.

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

Potential harms

- There is a possibility that this measure may inadvertently increase the overuse of bone mineral density tests and approved treatments for osteoporosis and fractures, especially in those who have a limited life expectancy. Although the population of women with recent osteoporotic fractures is least likely to be associated with overuse, the asymptomatic population is more prone to this. To help minimize this, the developer has an upper age limit of 85 for this measure and specific exclusions for those in hospice care and those living long-term in institutional settings.
- NCQA is also currently exploring additional exclusions to remove patients with advanced illness from this measure. These exclusions focus the measure on the population that is most likely to benefit from screening and treatment.

Additional Feedback: NA

***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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- The measure is publicly reported and used in current accountability programs.
- Public and payment reporting through CMS, Quality payment program, formerly PQRS
- It is selected as a reporting measure by 6.1 % of eligible providers."
- HEALTH PLAN LEVEL USE:
  - STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
  - CMS MEDICARE ADVANTAGE STAR RATING: This measure is included in the composite Medicare Advantage Star Rating.
  - HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
  - NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.

    - NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans.
    - NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- PHYSICIAN LEVEL USE
    - CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).
- Accountability and Feedback are satisfactorily addressed in the Measure Maintenance Document
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity.
- No additional information on use and reporting is provided in the 2018 submission.
- NCQA health plan rating process for health plans
- NCQA State of healthcare quality report annually
- NCQA health plan report cards
- NCQA ACO accreditation
- NCQA Health Plan accreditation
- NCQA Quality Compass
- CMS QPP for clinicians
- Publicly reported. Provider feedback available.

Usability

- The potential issues related to unintended harm have been addressed by the developer.
- Among physicians that reported on the measure to CMS, between 2009 and 2012 screening increased only 2.6 %.
- No identified harms. Results of various reporting programs (referenced below) may have a positive impact on quality/efficient health care.
- HEALTH PLAN LEVEL USE:
    - STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
    - CMS MEDICARE ADVANTAGE STAR RATING: This measure is included in the composite Medicare Advantage Star Rating.
    - HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
    - NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.
    - NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans.
    - NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- PHYSICIAN LEVEL USE
    - CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs)."
- Women who have had a fracture will have fewer fractures if osteoporosis is diagnosed and treated. The performance results should be acknowledged, publicized.
- No unintended consequences.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- No additional information on use for improvement is provided in the 2018 submission. There is mention in the original submission of the possibility of harms from overuse of bone density testing and osteoporosis treatments.

- No concerns;
- Health plans performance increased between 2014 and 2016
- Clinician performance increased by 13.5% but only .8% of clinicians reported on this measure
- No concerns; using top age of 85 so do not overuse in those with limited life expectancy;
- Benefits outweigh harms. Unintended consequences may be over treatment or use of second line agents for treatment of fragility fractures with a first-line treatment may be indicated.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0037 : Osteoporosis Testing in Older Women (OTO)
- 0046 : Screening for Osteoporosis for Women 65-85 Years of Age
- 2416 : Laboratory Investigation for Secondary Causes of Fracture
- 2417 : Risk Assessment/Treatment After Fracture

Harmonization

- There are multiple measures of osteoporosis prevention and management. During the last measure update in 2014, this measure was harmonized to align with applicable existing NQF-endorsed osteoporosis measures where possible given the different measure focus, methods of data collection and level of accountability. The developer provides a description of the harmonization between this measure (0053) and the most closely related measures, 0037, 0046, 2416, 2417.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0053

Measure Title: Osteoporosis Management in Women Who Had a Fracture

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

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

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

   ☒Yes (go to Question #2)

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

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

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

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

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☒High (go to Question #6)

   ☐Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.

   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☐Yes (go to Question #8)

   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

   REFERENCE: testing attachment section 2b1.

   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

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

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

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    REFERENCE: Testing attachment, section 2b2-2b6

    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

    ☒Yes (go to Question #12)

    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

    REFERENCE: Testing attachment, section 2b2.

    ***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the*

data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

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

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
    ☒Yes (go to Question #18)
    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
    ☒Yes (go to Question #19)
    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
    ☒Yes (go to Question #20)
    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #21)
    ☒Moderate (go to Question #21)
    ☐Low (please explain below then go to Question #21)
    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Prior validity studies of the same data elements may be submitted*
    ☒Yes (go to Question #22)
    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*
    ***REFERENCE:** Testing attachment, section 2b1.*
    ***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0053
Measure Title: Osteoporosis Management in Women Who Had a Fracture
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A
Date of Submission: 4/9/2018

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: Click here to name the health outcome

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

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Osteoporosis Management in Women Who Had a Fracture measures the percentage of women age 65 to 85 who receive a bone mineral density test or pharmacologic treatment for osteoporosis in the six months after a fracture.

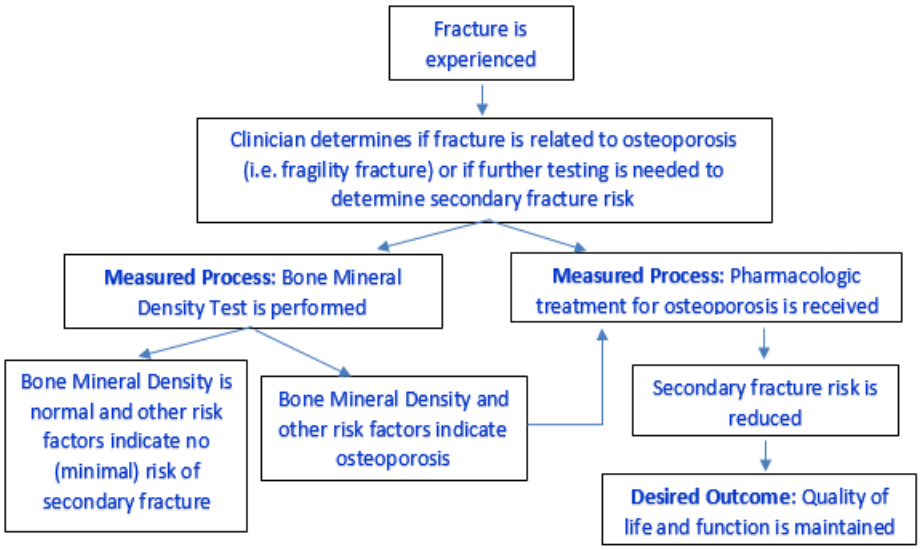
☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

2014 Submission:



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

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

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

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

Bone Mineral Density Testing After a Fracture

U.S. Preventive Services Task Force (USPSTF)	
Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2018 Submission** NCQA acknowledges that as of April 9, 2018, the U.S. Preventive Services Task Force (USPSTF) has released a DRAFT recommendation statement for osteoporosis screening. A draft Evidence Review was also published in November 2017. When published, NCQA will evaluate the final recommendation statement and supporting evidence review and consider any potential changes that may be needed for this measure. However, based on the draft recommendation statement we do not anticipate that any major revisions will be needed. U.S. Preventive Services Task Force. 2017. *Draft Recommendation Statement: Osteoporosis to Prevent Fractures: Screening*. https://www.uspreventiveservicestaskforce.org/Page/Document/draft-recommendation-statement/osteoporosis-screening1 U.S. Preventive Services Task Force. 2017. *Draft Evidence Review: Osteoporosis to Prevent Fractures: Screening.* https://www.uspreventiveservicestaskforce.org/Page/Document/draft-evidence-review/osteoporosis-screening1 **2014 Submission** U.S. Preventive Services Task Force. 2011. Screening for osteoporosis: US preventive services task force recommendation statement. Annals of internal medicine, 154(5), 356. http://www.uspreventiveservicestaskforce.org/uspstf10/osteoporosis/osteors.htm, accessed May 2, 2014.
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.	**2018 Submission** "The USPSTF recommends screening for osteoporosis with bone measurement testing to prevent osteoporotic fractures in women age 65 years and older. The USPSTF recommends screening for osteoporosis with bone measurement testing in postmenopausal women younger than age 65 years who are at increased risk of osteoporosis, as determined by a formal clinical risk assessment tool." – Experiencing a fracture is a significant factor in increasing fracture risk. **2014 Submission** "The USPSTF recommends screening for osteoporosis in women aged 65 years or older and in younger women whose fracture risk is equal to or greater than that of a 65-year-old white woman who has no additional risk factors." – Experiencing a fracture is a significant factor in increasing fracture risk.
Grade assigned to the **evidence** associated with the recommendation with the	**2018 Submission** The USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in women age 65 years and older is at least moderate. **2014 Submission** Moderate.

definition of the grade	
Provide all other grades and definitions from the evidence grading system	**2014 Submission** The USPSTF does not grade the evidence in the Evidence Based Practice report; they review the evidence and determine the certainty that there is benefit of an intervention. This certainty is based on the number, size and quality of individual studies but is not a grade of the evidence.
Grade assigned to the **recommendation** with definition of the grade	**2018 Submission** **Grade B:** The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. **2014 Submission** **Grade B**: The USPSTF recommends the services. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial Certainty Moderate: The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: • The number, size, or quality of individual studies • Inconsistency of findings across individual studies • Limited generalizability of findings to routine primary care practice • Lack of coherence in the chain of evidence
Provide **all other grades and definitions** from the recommendation grading system	**2018 Submission** Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. 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. Grade I: 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. **2014 Submission** **Grade A**: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Certainty High: The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.

	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. Certainty Low: The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: • The limited number or size of studies • Important flaws in study design or methods • Inconsistency of findings across individual studies • Gaps in the chain of evidence • Findings not generalizable to routine primary care practice • Lack of information on important health outcomes
Body of evidence: • **Quantity** – how many studies? • **Quality** – what type of studies?	**2018 Submission** The DRAFT evidence report (Viswanathan et al 2017) supporting this guideline outlines the quantity and quality of evidence, which are summarized below for the key questions of the 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? • As in the previous 2011 review, found no good or fair quality studies eligible for this key question Key Question 2a. What is the accuracy and reliability of screening approaches to identify adults who are at increased risk for osteoporotic fracture? • Accuracy of Clinical Risk Assessment Tools for Identifying Osteoporosis: included 37 articles (35 studies, fair or good quality) • Accuracy of Bone Measurement Tests Used to Identify Low Bone Mass and Osteoporosis: included 11 studies, fair or good quality • Accuracy of Bone Measurement Tests Used to Predict Fracture: included 21 studies, fair or good quality • Accuracy of Fracture Risk Prediction Instruments: included 1 systematic review and 13 fair or good quality observational studies Key Question 2b. What is the evidence to determine screening intervals and how do these vary by baseline fracture risk? • Included 2 articles (2 studies, good quality) Key Question 3. What are the harms of screening for osteoporotic fracture risk? • Found no eligible studies that addressed this question Key Question 4a. What is the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality? Bisphosphonates: • Alendronate: included 7 studies, fair or good quality • Zoledronic Acid: included 2 studies, fair or good quality • Risedronate: included 4 studies, fair or good quality

	• Etidronate: included 2 fair quality studies • Ibandronate: identified no studies or trials that assessed the benefits of ibandronate for preventing fractures Raloxifene: • Included 1 large good quality RCT Estrogen: • No studies included Denosumab: • Included 3 fair quality trials Parathyroid Hormone: • Included 2 fair quality trials Key Question 4b. How does the effectiveness of pharmacotherapy for the reduction of fractures and related morbidity and mortality vary by subgroup, specifically in postmenopausal women, premenopausal women, men, younger age groups (age <65 years), older age groups (age ≥65 years), baseline bone mineral density, and baseline fracture risk? Bisphosphonates: • Zoledronic Acid, Etidronate, Ibandronate: found no relevant results in included studies for subgroup analysis • Alendronate: included 1 study • Risedronate: included 1 RCT Raloxifene: • Included 1 study Estrogen: • No studies included Denosumab: • Included 1 fair quality trial Parathyroid Hormone: • Included 1 fair quality trial Key Question 5. What are the harms associated with pharmacotherapy? Bisphosphonates: • Alendronate: included 16 studies, fair or good quality • Zoledronic Acid: included 4 studies, fair or good quality • Risedronate: included 4 studies, fair or good quality • Etidronate: included 2 fair quality studies • Ibandronate: included 7 fair quality studies Raloxifene: • Included 6 studies Estrogen: • No studies included Denosumab: • Included 3 fair quality studies Parathyroid Hormone: • Included 2 fair quality studies Viswanathan, M., et al. 2017. “Screening to Prevent Osteoporotic Fractures: An Evidence Review for the U.S. Preventive Services Task Force.” Available here: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryDraft/osteoporosis-screening1

	2014 Submission N/A – not required in previous submission
Estimates of **benefit and consistency** across studies	**2018 Submission** The following text is quoted directly from the USPSTF recommendation statement. The USPSTF found no studies that evaluated the effect of screening for osteoporosis on fracture rates or fracture-related morbidity or mortality. The USPSTF found convincing evidence that bone measurement tests are accurate for detecting osteoporosis and predicting osteoporotic fractures in women and men. The USPSTF found adequate evidence that clinical risk assessment tools are moderately accurate in identifying risk of osteoporosis and osteoporotic fractures. The USPSTF found convincing evidence that drug therapies reduce subsequent fracture rates in postmenopausal women. The benefit of treating screening-detected osteoporosis is at least moderate in women age 65 years and older and younger postmenopausal women who have similar fracture risk. The harms of treatment range from no greater than small for bisphosphonates and parathyroid hormone to small to moderate for raloxifene and estrogen. Therefore, the USPSTF concludes with moderate certainty that the net benefit of screening for osteoporosis in these groups of women is at least moderate. The USPSTF concludes that the evidence is inadequate to assess the effectiveness of drug therapies in reducing subsequent fracture rates in men without previous fractures. Treatments that have been proven effective in women cannot necessarily be presumed to have similar effectiveness in men, and the direct evidence is too limited to draw definitive conclusions. Thus, the USPSTF could not assess the balance of benefits and harms of screening for osteoporosis in men. **2014 Submission** N/A
What harms were identified?	**2018 Submission** The following is quoted directly from the USPSTF draft recommendation statement: "The USPSTF found no studies that described harms of screening for osteoporosis in men or women. Based on the nature of screening with bone measurement tests and the low likelihood of serious harms, the USPSTF found adequate evidence to bound these harms as no greater than small. Harms associated with screening may include radiation exposure from DXA and opportunity costs (time and effort required by patients and the health care system)." **2014 Submission** Potential harms of bone mineral density testing: The USPSTF found no new studies that described harms of screening for osteoporosis in men or women. Screening with DXA is associated with opportunity costs (time and effort required by patients and the health care system). Potential harms of screening for osteoporosis include false-positive test results causing unnecessary treatment, false-negative test results, and patient anxiety about positive test results (USPSTF 2011). The USPSTF

		concluded that there is moderate certainty that the net benefit of screening for osteoporosis by using DXA is at least moderate.
Bone	Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2018 Submission** To our knowledge, there have been no published studies since the systematic review that would impact the recommendations above. When the USPSTF final evidence review is published, NCQA will conduct further review to determine if there are any changes to the evidence that would warrant refinements to the measure.

Mineral Density Testing after Fracture

American Association of Clinical Endocrinologists (AACE)	
Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2018 Submission** Screening women who had a fragility fracture. AACE (2016) **American Association of Clinical Endocrinologists (AACE).** Clinical Practice Guidelines for the Diagnosis and Treatment of Postmenopausal Osteoporosis, 2016, Sep. Guideline available from https://www.aace.com/files/postmenopausal-guidelines.pdf. **2014 Submission** Screening women who had a fragility fracture. AACE (2010) **American Association of Clinical Endocrinologists (AACE).** Medical Guidelines for Clinical Practice for the Diagnosis and Treatment of Postmenopausal Osteoporosis, 2010 Dec. Guideline available from https://www.aace.com/files/osteo-guidelines-2010.pdf, accessed April 25, 2014.
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.	**2018 Submission** "Who needs to be screened for osteoporosis? All postmenopausal women ≥ 50 years should undergo clinical assessment for osteoporosis and fracture risk, including a detailed history and physical examination." (page 7) "The AACE recommends bone mineral density testing for women aged 65 and older and younger postmenopausal women at increased risk for bone loss and fracture based on fracture risk analysis." (page 10) **2014 Submission** Who needs to be screened for osteoporosis? Younger postmenopausal women at increased risk of fracture, based on a list of risk factors - "Indications for bone mineral testing: All postmenopausal women with a history of fracture without major trauma after age 40 to 45." (page 17)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2018 Submission** Level 2. **2014 Submission** Level 2.
Provide all other grades and definitions from the evidence grading system	**2018 Submission** 2016 AACE Guidelines used the 2010 Criteria for Rating of Published Evidence Table submitted in 2014. **2014 Submission**

Table 2: 2010 American Association of Clinical Endocrinologists Criteria for Rating of Published Evidence. 1 = strong evidence; 2 = intermediate evidence; 3 = weak evidence; 4 = no evidence.

Numerical descriptor (evidence level)	Study Type
1	Meta-analysis of randomized controlled trials
1	Randomized controlled trial
2	Meta-analysis of nonrandomized prospective or case-controlled trials
2	Nonrandomized controlled trial
2	Prospective cohort study
2	Retrospective case-control study
3	Cross-sectional study
3	Surveillance study (registries, surveys, epidemiologic study)
3	Consecutive case series
3	Single case reports
4	No evidence (theory, opinion, consensus, or review

Grade assigned to the **recommendation with definition** of the grade

2018 Submission

Grade B.

- Evidence from at least 1 large well-designed clinical trial, cohort or case-controlled analytic study, or meta-analysis.
- No conclusive level 1 publication; ≥ 1 conclusive level 2 publications demonstrating benefit > risk. (see Table 2 in section above)

2014 Submission

Grade C.

- Evidence based on clinical experience, descriptive studies, or expert consensus opinion.
- No conclusive level 1 or 2 publications; ≥ 1conclusive level 3 publications demonstrating benefit > risk.
- No conclusive risk at all and no conclusive benefit demonstrated by evidence. (see Table 2)

Provide all other grades and definitions from the recommendation grading system

2018 Submission

AACE Grade Definition

Grade A:

- Homogeneous evidence from multiple well-designed randomized controlled trials with sufficient statistical power.
- Homogenous evidence from multiple well-designed randomized or cohort controlled trials with sufficient statistical power.
- ≥ 1 conclusive level 1 publications demonstrating benefit > risk. (see Table 2)

Grade C:

- Evidence based on clinical experience, descriptive studies, or expert consensus opinion.
- No conclusive level 1 or 2 publications; ≥ 1conclusive level 3 publications demonstrating benefit > risk.
- No conclusive risk at all and no conclusive benefit demonstrated by evidence. (see Table 2)

Grade D:

- Not rated.
- No conclusive level 1, 2, or 3 publication demonstrating benefit > risk.
- Conclusive level 1, 2, or 3 publication demonstrating risk > benefit. (see table 2)

2014 Submission

AACE Grade Definition

Grade A:

	• Homogeneous evidence from multiple well-designed randomized controlled trials with sufficient statistical power. • Homogenous evidence from multiple well-designed randomized or cohort controlled trials with sufficient statistical power. • ≥ 1 conclusive level 1 publications demonstrating benefit > risk. (see Table 2) **Grade B**: • Evidence from at least 1 large well-designed clinical trial, cohort or case-controlled analytic study, or meta-analysis. • No conclusive level 1 publication; ≥ 1 conclusive level 2 publications demonstrating benefit > risk. (see Table 2) **Grade D**: • Not rated. • No conclusive level 1, 2, or 3 publication demonstrating benefit > risk. • Conclusive level 1, 2, or 3 publication demonstrating risk > benefit. (see table 2)
Body of evidence: • **Quantity** – how many studies? • **Quality** – what type of studies?	**2014 Submission** Although the AACE guidelines above were based on a systematic evidence reviews, they did not provide a summary of the evidence (quantity, quality and consistently) to answer the questions laid out in the NQF submission for this measure. Therefore, NCQA supplemented the guidelines with the systematic reviews documented below.
Estimates of **benefit and consistency** across studies	**2014 Submission** N/A
What **harms** were identified?	**2014 Submission** N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2018 Submission** To our knowledge, there have been no published studies since the systematic review that would impact the recommendations above.

Bone Mineral Density Testing after Fracture

Other Systematic Review of the Body of Evidence	
Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2014 Submission** Although the AACE guidelines above were based on systematic evidence reviews, they did not provide a summary of the evidence (quantity, quality and consistently) to answer the questions laid out in the NQF submission for this measure. Therefore, we supplemented the guidelines with the following systematic reviews. Nelson, H. D., Haney, E. M., Chou, R., Dana, T., Fu, R., & Bougatsos, C. (2010). Screening for Osteoporosis. Systematic Review to Update the 2002 U.S. Preventive Services Task Force Recommendation. Rockville (MD): Agency for Healthcare Research and Quality (US). Crandall, C. J., Newberry, S. J., Diamant, A., Lim, Y. W., Gellad, W. F., Suttorp, M. J., ... & Shekelle, P. G. (2012). Treatment to prevent fractures in men and women with low bone density or osteoporosis: update of a 2007 report. Rockville (MD): Agency for Healthcare Research and Quality (US)

	Levis, S., & Theodore, G. (2012). Summary of AHRQ's comparative effectiveness review of treatment to prevent fractures in men and women with low bone density or osteoporosis: update of the 2007 report. *Journal of managed care pharmacy: JMCP*, *18*(4 Suppl B), S1.
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.	**2014 Submission** This measure assesses whether a patient who had a fracture of any bone other than face, finger, toe, or skull was given the appropriate follow-up care post-fracture to prevent a secondary fracture from occurring. Appropriate follow-up care includes either 1) bone mineral density testing to assess whether a patient has osteoporosis or 2) receiving pharmacologic therapy to treat osteoporosis. This measure is based on guidelines and evidence that patients at high risk of fracture, including patients with a history of fragility fractures, should be screened for osteoporosis (USPSTF 2012, Nelson 2010, AACE 2010) and that patients who have a fragility fracture of the hip or spine should be provided with a treatment for osteoporosis (AACE 2010, Crandall 2012). AACE defines a fragility fracture as "a fracture that results from trauma less than or equal to that from a fall from a standing height and almost always indicates decreased skeletal strength." Administrative claims (one of the data sources for this measure) cannot determine if a fracture meets this definition of a fragility fracture. Therefore, we remove fractures from the measure that are rarely fragility related (face, finger, toe or skull) and designed the measure to allow the provider the flexibility to determine if either a bone mineral density test or pharmacologic treatment is the best follow-up intervention for post-menopausal women who experienced a fracture. If a fragility fracture can be assumed the provider can meet the measure numerator criteria by providing the appropriate drug therapy. If a fragility fracture cannot be assumed (i.e. the fracture was associated with trauma such as a car accident) the provider may choose to screen for underlying osteoporosis risk as recommended for all women who experience a fracture. This measure design reduces the possibility of overtreatment of osteoporosis in women who experience fractures.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2014 Submission** N/A
Provide **all other grades and definitions from the evidence** grading system	**2014 Submission** N/A

Grade assigned to the **recommendation** with definition of the grade	**2014 Submission** N/A
Provide all **other grades and definitions from the recommendation** grading system	**2014 Submission** N/A
Body of evidence: • **Quantity** – how many studies? • **Quality** – what type of studies?	**2014 Submission** **Quantity** Nelson (2012): This update to the evidence review did not include a key question addressing the efficacy of bone mineral density tests to predict fracture risk. This was established in the 2002 Evidence Review and not re-evaluated in 2012. The evidence review did include a key question of pertinence to this measure. Key Question 3b: How well do peripheral bone measurement tests predict fractures? Six prospective studies and one meta-analysis comparing DXA to Quantitative Ultrasonography were used to address this question. **Quality** Nelson (2012): How well do peripheral bone measurement tests predict fractures? The authors described the six prospective studies as large and well-designed and do not note any quality concerns. The meta-analysis used to compare DXA with Qualitative Ultrasonography (QUS) included multiple studies that varied by subject characteristics including location (Europe, United States, Asia), sample size (110-722), prevalence of osteoporosis (7-38 percent), age (46-64 years), and sex. No studies described race or ethnicity of subjects. Potential sources of bias included insufficient information to determine participant selection methods, time between QUS and DXA, and whether QUS and DXA results were interpreted independently of each other.
Estimates of benefit and consistency across studies	**2014 Submission** Nelson (2012): How well do peripheral bone measurement tests predict fractures? "Several peripheral bone measurement tests have been developed, although clinical practice and recent research focus on QUS of the calcaneous (heel). Large studies of postmenopausal women and men indicate that QUS obtained at the calcaneus using various types of devices can predict fractures as well as DXA of the femoral neck, hip, or spine, although variation exists across studies. However, QUS is not a good predictor of DXA as determined by a recent meta-analysis that indicated AUC estimates of 0.74–0.77 depending on the QUS parameter used. Also, it is unclear how results of QUS can be used to select individuals for drug therapies that were proven efficacious based on DXA criteria." "Overall, DXA and QUS have similar area under the curve (AUC) estimates and odds ratios for fracture outcomes (Table 4). For all fractures combined, AUC estimates range from 0.59–0.66 and ORs from 1.81–2.16 for DXA of the femoral neck. For QUS, AUC estimates are approximately 0.60, and ORs range from 1.26–2.25. In one study that included DXA of the distal radius, the AUC estimate was 0.64 (95% CI, 0.59–0.68) and OR for all fractures 1.47 (95% CI, 1.28–1.68). "QUS predicts most fractures as well as DXA and offers distinct advantages, such as lower cost, portability, ease of use, and avoidance of ionizing radiation. However, it is not clear how to apply the results of QUS testing to patient management. Currently, standardized diagnostic criteria for osteoporosis uses DXA not QUS cutpoints, and clinical trials of drug therapies used DXA testing in

	its selection criteria. To be clinically useful, QUS results would need to be similar to DXA."
What harms were identified?	**2014 Submission** N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2014 Submission** No.

Pharmacologic Therapy After a Fracture

American Association of Clinical Endocrinologists (AACE)	
Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2018 Submission** **American Association of Clinical Endocrinologists (AACE).** Clinical Practice Guidelines for the Diagnosis and Treatment of Postmenopausal Osteoporosis, 2016, Sep. Guideline available from https://www.aace.com/files/postmenopausal-guidelines.pdf. **2014 Submission** **American Association of Clinical Endocrinologists (AACE).** Medical Guidelines for Clinical Practice for the Diagnosis and Treatment of Postmenopausal Osteoporosis, 2010 Dec. Guideline available from https://www.aace.com/files/osteo-guidelines-2010.pdf, accessed April 25, 2014.
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.	**2018 Submission** "Strongly recommend pharmacologic therapy for patients with osteopenia or low bone mass and a history of fragility fracture of the hip or spine." (page 4) **2014 Submission** "Patients who have a history of fracture of the hip or spine need pharmacologic therapy." (Page 4)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2018 Submission** Level 1. **2014 Submission** Level 1.
Provide all other grades and definitions from the evidence grading system	**2018 Submission** 2016 AACE Guidelines used the 2010 Criteria for Rating of Published Evidence Table submitted in 2014. **2014 Submission**

Table 2: 2010 American Association of Clinical Endocrinologists Criteria for Rating of Published Evidence. 1 = strong evidence; 2 = intermediate evidence; 3 = weak evidence; 4 = no evidence.

Numerical descriptor (evidence level)	Study Type
1	Meta-analysis of randomized controlled trials
1	Randomized controlled trial
2	Meta-analysis of nonrandomized prospective or case-controlled trials
2	Nonrandomized controlled trial
2	Prospective cohort study
2	Retrospective case-control study
3	Cross-sectional study
3	Surveillance study (registries, surveys, epidemiologic study)
3	Consecutive case series
3	Single case reports
4	No evidence (theory, opinion, consensus, or review

Grade assigned to the **recommendation with definition** of the grade

2018 Submission

Grade A

- Homogeneous evidence from multiple well-designed randomized controlled trials with sufficient statistical power.
- Homogenous evidence from multiple well-designed randomized or cohort controlled trials with sufficient statistical power.
- ≥ 1 conclusive level 1 publications demonstrating benefit > risk. (see Table 2 in previous section)

2014 Submission

Grade A

- Homogeneous evidence from multiple well-designed randomized controlled trials with sufficient statistical power.
- Homogenous evidence from multiple well-designed randomized or cohort controlled trials with sufficient statistical power.
- ≥ 1 conclusive level 1 publications demonstrating benefit > risk. (see Table 2 in previous section)

Provide **all other grades and definitions** from the recommendation grading system

2018 Submission

2016 AACE Guidelines used the 2010 AACE Protocol for Production of Clinical Practice Guidelines submitted in 2014.

2014 Submission

AACE Grade Definition

Grade B:

- Evidence from at least 1 large well-designed clinical trial, cohort or case-controlled analytic study, or meta-analysis.
- No conclusive level 1 publication; ≥ 1 conclusive level 2 publications demonstrating benefits > risk. (see Table 2)

Grade C:

- Evidence based on clinical experience, descriptive studies, or expert consensus opinion.
- No conclusive level 1 or 2 publications; ≥ 1conclusive level 3 publications demonstrating benefit > risk.
- No conclusive risk at all and no conclusive benefit demonstrated by evidence. (see Table 2)

Grade D:

- Not rated.
- No conclusive level 1, 2, or 3 publication demonstrating benefit > risk.
- Conclusive level 1, 2, or publication demonstrating risk > benefit. (see Table 2)

Body of evidence: • **Quantity** – how many studies? • **Quality** – what type of studies?	**2014 Submission** Although the AACE guidelines above were based on systematic evidence reviews, they did not provide a summary of the evidence (quantity, quality and consistently) to answer the questions laid out in the NQF submission for this measure. Therefore, NCQA supplemented the guidelines with the systematic reviews below.
Estimates of benefit and consistency across studies	**2014 Submission** N/A
What **harms** were identified?	**2014 Submission** Potential harms of pharmacologic treatment for osteoporosis: The AACE guideline and evidence above outlines the potential harms and side effects related to each pharmacologic treatment for osteoporosis: • Bisphosphonates: The most common side effect from bisphosphonates is esophageal irritation. • Calcitonin: Nausea, local inflammatory reactions at the injection site, and vasomotor symptoms including sweating and flushing. • Teriparatide: Nausea, orthostatic hypotension, and leg cramps. Hypercalcemia has been observed but is not common. • Demosumab: Before initiation of therapy, hypocalcemia must be corrected. Serious infections such as skin or cellulitis can occur. Dermatitis, rashes, eczema and osteonecrosis of the jaw has been reported. • Raloxifene: Associated with an approximate three-fold increase in occurrence of venous thromboembolic diseases, menopausal symptoms (hot flashes) and leg cramps.
Identify any **new studies conducted since the SR**. Do the new studies change the conclusions from the SR?	**2014 Submission** No.

Pharmacologic Therapy After a Fracture

Other Systematic Review of the Body of Evidence	
Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2014 Submission** Although the AACE guidelines above were based on systematic evidence reviews, they did not provide a summary of the evidence (quantity, quality and consistently) to answer the questions laid out in the NQF submission for this measure. Therefore, we supplemented the guidelines with the following systematic reviews. Nelson, H. D., Haney, E. M., Chou, R., Dana, T., Fu, R., & Bougatsos, C. (2010). Screening for Osteoporosis. Systematic Review to Update the 2002 U.S. Preventive Services Task Force Recommendation. Rockville (MD): Agency for Healthcare Research and Quality (US). Crandall, C. J., Newberry, S. J., Diamant, A., Lim, Y. W., Gellad, W. F., Suttorp, M. J., ... & Shekelle, P. G. (2012). Treatment to prevent fractures in men and women with low bone density or osteoporosis: update of a 2007 report. Rockville (MD): Agency for Healthcare Research and Quality (US) Levis, S., & Theodore, G. (2012). Summary of AHRQ's comparative effectiveness review of treatment to prevent fractures in men and women with low bone density or osteoporosis: update of the 2007 report. *Journal of managed care pharmacy: JMCP*, *18*(4 Suppl B), S1.

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.	**2014 Submission** N/A
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2014 Submission** The only evidence review above with graded evidence was Crandall (2012). The table below provides a summary of the evidence grades presented in Crandall (2012). **TABLE 3** Strength of Evidence for the Reduction of Risk of Fracture Types with Pharmacotherapy in Women with Postmenopausal Osteoporosis Fracture Skeletal Sites Agent / Vertebral / Nonvertebral / Hip / Wrist Alendronate ●●● ●●● ●●● ● Ibandronate ●●● ●● ● I Risedronate ●●● ●●● ●●● ● Zoledronic acid ●●● ●●● ●●● I Denosumab ●●● ●●● ●●● I Teriparatide ●●● ●● ● I Raloxifene ●●● I I I *Source: Table A in: Crandall CC, Newberry SJ, Gellad WG et al. Comparative effectiveness of treatment to prevent fractures in men and women with low bone density or osteoporosis - update of a 2007 report. AHRQ comparative effectiveness review no. 53. March 2012.*[6] *Strength of evidence symbol legend: I = insufficient strength of evidence; ● = low strength of evidence; ●● = moderate strength of evidence; ●●● = high strength of evidence.*
Provide all **other grades and definitions from the evidence** grading system	**2014 Submission** Shown in Table above.
Grade assigned to the **recommendation** with definition of the grade	**2014 Submission** N/A
Provide **all other grades and definitions from the recommendation** grading system	**2014 Submission** N/A
Body of evidence: • **Quantity** – how many studies? • **Quality** – what type of studies?	**2014 Submission** **Quantity** Crandall 2012: Key Question 1 What Are the Comparative Benefits in Fracture Risk Reduction Among the Following Therapeutic Modalities for low Bone Density: Bisphosphonates, Denosumab, Menopausal Hormone Therapy, Selective Estrogen Receptor Modulators (Raloxifene), Parathyroid Hormone, Calcium, Vitamin D, and Physical Activity? "For this question, we identified 55 RCTs and 10 observational studies in addition to 58 systematic reviews (from both the original and current report) that assessed the effects of interventions compared to placebo: nine systematic reviews and 10 RCTs for alendronate, 10 systematic reviews and 13 RCTs for risedronate, three systematic reviews and three RCTs for ibandronate, four RCTs for zoledronic acid, one systematic review and two RCTs for denosumab, three systematic review and three RCTs for raloxifene, two systematic reviews

	and three RCTs for teriparatide, six RCTs for menopausal estrogen therapy, four systematic reviews and six RCTs for calcium alone, 15 systematic reviews and seven RCTs for vitamin D alone, four RCTs for vitamin D plus calcium, and one systematic review and one RCT for physical activity." **Quality** Crandall (2012): Comparative Effectiveness of Treatments: The authors rated the quality of evidence for each study but did not discuss any specific concerns about the quality of evidence or sources of bias. Overall, they found the majority of evidence came from well-designed large RCTs.
Estimates of **benefit and consistency** across studies	**2014 Submission** Overall there is moderate certainty that bone mineral density tests predict future fracture risk and pharmacologic treatment for individuals at risk of future fracture reduces the fracture risk. Crandall (2012): Comparative Effectiveness of Treatments There were many different pharmacologic treatments that were determined effective at reducing future fractures in patients who are at high risk. The review concluded that Alendronate, etidronate, ibandronate, risedronate, teriparatide, denosumab, and raloxifene reduce the risk of fractures among high risk groups including postmenopausal women with osteoporosis. For the purposes of this review "High Risk" was defined as the following: 1) transplant population, or 2) study entry criteria require T score ≤ -2.5, or 3) study entry criteria require≥1 fracture, or 4) ≥50% of population has 1 or more fractures at baseline, or 5) Significant neuromuscular impairment The table below shows selected studies from the review by Crandall et al. and previous review *by McLean et al. covering 1996-2006.* These studies indicate decreased odds of fracture with medicaiton compared to placebo or control group among high/intermediate risk populations. **Table 3**: Risk of fracture for medication relative to placebo – Selected studies from Crandall et al. 2012 **SEE APPENDIX A for Table Information**
What **harms** were identified?	Potential harms of pharmacologic treatment for osteoporosis: The AACE guideline **and evidence above outlines** the potential harms and side effects related to each pharmacologic treatment for osteoporosis: • Bisphosphonates: The most common side effect from bisphosphonates is esophageal irritation. • Calcitonin: Nausea, local inflammatory reactions at the injection site, and vasomotor symptoms including sweating and flushing. • Teriparatide: Nausea, orthostatic hypotension, and leg cramps. Hypercalcemia has been observed but is not common. • Demosumab: Before initiation of therapy, hypocalcemia must be corrected. Serious infections such as skin or cellulitis can occur. Dermatitis, rashes, eczema and osteonecrosis of the jaw has been reported. • Raloxifene: Associated with an approximate three-fold increase in occurrence of venous thromboembolic diseases, menopausal symptoms (hot flashes) and leg cramps.
Identify any **new studies conducted since the SR**. Do the new studies	**2014 Submission** Eriksen et al. (2014) conducted a systematic review on the use of long-term treatment with bisphosphonates for postmenopausal osteoporosis. This

change the conclusions from the SR?	review found that long-term use of bisphosphonates resulted in fewer fractures and smaller loss of bone mineral density in women who remained on treatment for three or more years. Residual benefits were found for women who received alendronate or zoledronic acid as long as they received initial treatment of 3-5 years. Residual benefits were seen even after they discontinued treatment for 3-5 years. Overall, this review found "BMD monitoring and fracture risk assessments should be conducted regularly to determine whether treatment could be stopped or if it should be reinitiated." Eriksen EF, Díez-Pérez A, Boonen S. Update on long-term treatment with bisphosphonates for postmenopausal osteoporosis: a systematic review. Bone. 2014 Jan;58:126-35. doi: 10.1016/j.bone.2013.09.023. Available at http://www.thebonejournal.com/article/S8756-3282(13)00378-5/abstract

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

APPENDIX A

Table 3: Risk of fracture for medication relative to placebo – Selected studies from Crandall et al. 2012

Author, year	Study duration	Type of fracture	Risk level*	# of fractures, medication	# of fractures, placebo or control†	Odds ratio (95% CI)
Alendronate						
Quandt, 2005	54 months	Vertebral fractures	Intermediate	12/1878	29/1859	0.43(0.23, 0.79)
Ibandronate						
Chesnut, 2004	36 months	clinical vertebral	High	44/1954	41/975	0.50 (0.32, 0.79)
Risendronate						
Sato, 2005	18 months	nonvertebral fracture	High	8/231	29/230	0.29 (0.15, 0.57)
Sato, 2005	18 months	hip fracture	High	5/231	19/230	0.29 (0.13, 0.66)
Zoledronic acid (5 milligrams once)						
Black, 2007	36 months	Any clinical; fracture	High	308/3667	456/3563	0.63 (0.54, 0.72)
Calcitonin						
Ishida, 2004 20 IU weekly	24 months	vertebral	High	8/66	17/66	0.41 (0.17, 0.99)
Toth, 2005 200 IU daily, alternate months	18 months	vertebral fracture	High	0/40	3/31	0.09 (0.01, 0.96)
Teriparatide						
Gallagher, 2005	21 months	vertebral fracture	High	22/403	62/398	0.34 (0.22, 0.54)
Raloxifene						
Barrett-Connor, 2006	5.6 years	Clinical vertebral	Unknown risk	64/5,044	97/5,057	0.66 (0.48. 0.90)
Denosumab						
Cummings, 2009	36 months	Hip fracture	Unknown risk	26/3,714	43/3,583	0.59 (0.36, 0.94)
Cummings, 2009	36 months	Nonvertebral	Unknown risk	238/3,662	293/3,663	0.8 (0.67, 0.95)

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0053
Corresponding Measures:
De.2. **Measure Title:** Osteoporosis Management in Women Who Had a Fracture
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of women age 50-85 who suffered a fracture and who either had a bone mineral density test or received a prescription for a drug to treat osteoporosis.
1b.1. **Developer Rationale:** The intent of this measure is secondary prevention of fractures through the appropriate diagnosis and treatment of osteoporosis. Detecting osteoporosis and initiating treatment will help to prevent future fractures from occurring. Future fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

S.4. **Numerator Statement:** Patients who received either a bone mineral density test or a prescription for a drug to treat osteoporosis after a fracture occurs.
S.6. **Denominator Statement:** Women who experienced a fracture, except fractures of the finger, toe, face or skull. Three denominator age strata are reported for this measure:
Women age 50-64
Women age 65-85
Women age 50-85
S.8. **Denominator Exclusions:** - Exclude women who had a bone mineral density test during the 24 months prior to the index fracture.
- Exclude women who had a claim/encounter for osteoporosis treatment during 12 months prior to the index fracture.
- Exclude women who received a dispensed prescription or had an active prescription to treat osteoporosis during the 12 months prior to the index fracture.
- Exclude women who are enrolled in a Medicare Institutional Special Needs Plan (I-SNP) or living long-term in an institution any time during the measurement year.
- Exclude women receiving hospice care during the measurement year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records
S.20. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Dec 30, 2014

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

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0053_OMW_Evidence_FINAL.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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)*
<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
The intent of this measure is secondary prevention of fractures through the appropriate diagnosis and treatment of osteoporosis. Detecting osteoporosis and initiating treatment will help to prevent future fractures from occurring. Future fractures, especially in the older population, can cause significant health issues, decline in function, and, in some cases lead to mortality.

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis. *(<u>This is required for maintenance of endorsement</u>. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
HEALTH PLAN LEVEL:
Performance Rates: The following data are extracted from HEDIS data collection for Medicare Advantage Health Plans and reflect the most recent years of measurement for this measure. Performance data are summarized at the health plan level and described by mean, standard deviation, and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year.

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH
2014 | 35.9% | 17.3% | 15.8% | 22.6% | 33.7% | 45.9% | 58.0%
2015 | 38.7% | 17.9% | 17.6% | 24.1% | 36.4% | 49.0% | 75.51%
2016 | 40.0% | 19.0% | 17.4% | 24.6% | 38.6% | 51.7% | 76.4%

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans reporting the HEDIS measure and the mean eligible population for the measure across these health plans.

Year | N Plans | Avg Eligible Population per Plan | SD
2014 | 302 | 570.9 | 469.1
2015 | 279 | 824.5| 384.4
2016 | 277 | 817.0 | 402.3

PHYSICIAN LEVEL:
The following data are extracted from Physician Quality Reporting System (PQRS) and reflect claims data for services provided from January 1, 2009 through December 31, 2011 . PQRS refers to the pay-for-reporting incentive program that allowed providers to choose which quality measures to report on. As of 2017, PQRS has been renamed as QPP, the Quality Payment Program. In 2012, of 204,369 eligible providers, only 0.8% chose to report on this measure. Therefore, the performance rates below are reflective of less than one percent of Medicare providers. At the time of data collection this measure applied to women age 50 and older. In 2014 the measure was revised to reflect the added upper age limit. For the next year of quality measurement reporting, the physician level performance will be reported for the 50-85 age strata. This strata was selected for reporting because it is the broadest age range.

Performance data is summarized at the physician level and described by mean, 10th, 25th, 50th, 75th and 90th percentile.

Performance Rate for all Reporting Providers for 2012
Mean|10th| 25th| 50th | 75th | 90th

70.0%|0.00% | 25.0% | 100% | 100% | 100%

The following data (also extracted from PQRS) show the average performance rates for several years prior to 2012.

Average performance rates from 2009-2011
2009 | 56.5%
2010 | 46.8%
2011 | 70.6%

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.
N/A

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.*
Health Plan Reporting:
NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escarce et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escarce, 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership were designed to promote standardized methods for collecting these data. These measures follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

Escarce JJ, Carreón R, Veselovskiy G, Lawson EH. Collection of race and ethnicity data by health plans has grown substantially, but opportunities remain to expand efforts. Health Aff (Millwood). 2011;30(10):1984-1991. - See more at: http://www.ajmc.com/publications/issue/2012/2012-7-vol18-n7/exploring-health-plan-perspectives-in-collecting-and-using-data-on-race-ethnicity-and-language/4#sthash.23sL3Iuc.dpuf

Physician Level Reporting:
CMS does not currently report performance data stratified by different variables in the PQRS/QPP program, where the measure is in use.

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
People of all ethnic backgrounds are at risk of osteoporosis; however, non-Hispanic Caucasian and Asian women 50 and older have a higher prevalence of osteoporosis (20 percent), compared with Hispanic (10 percent) and non-Hispanic African American (5 percent) populations (NOF, 2013). Similarly, hip fracture rates are highest for non-Hispanic Caucasian women (140.7 per 100,000) and Asian women (85.4 per 100,000), but still prevalent in African American women (57.3 per 100,000) and Hispanic women (49.7 per 100,000) (Silverman, 1988).

Research suggests that African American women receive less dual-energy x-ray absorptiometry screenings and treatment for osteoporosis. One study found that 30% (21% received test) of African American women were referred to dual-energy x-ray absorptiometry tests compared to 38% (27% received test) of Caucasian women. In addition, for those women who had a confirmed diagnosis of osteoporosis, 78% of African American women were likely to receive a medication compared to 89% of

Caucasians (Hamrick, 2012). An earlier study with a smaller sample size found that of those diagnosed with osteoporosis, 62% of African Americans were started on a treatment compared to 83% of Caucasian women (Hamrick, 2006).

In a cohort study of patients identified by the Indiana Health Information Exchange, African American women had the lowest treatment rates for osteoporosis when compared with women of other races. The cohort was comprised of 36,965 patients (10.7% African Americans, 81.3% non-African American, 8.1% unreported) between 2005 and 2011 with at least one osteoporotic event (Liu, 2016). Of the 3,943 African-American women enrolled in the study, 17.6% began treatment within 2 years of the index event compared with 23.7% for non-African American women (p value <.0001) (Liu, 2016). Overall, 23.3% of all patients identified in this cohort received treatment within the 2 years following the index event (Liu, 2016).

These studies highlight an opportunity to improve screening and timely treatment for all individuals with osteoporotic events, but particularly for African American women.

Hamrick I, Cao Q, Agbafe-Mosley D, Cummings DM. Osteoporosis healthcare disparities in postmenopausal women. J Womens Health (Larchmt). 2012 Dec; 21 (12):1232-6. Doi: 10.1089/jwh.2012.3812. Epub 2012 Nov 9.

Hamrick, I, Whetsone LM, Cummings DM. Racial disparity in treatment of osteoporosis after diagnosis. Osteoporos Int. 2006;17 (11): 1653-8. Epub 2006 Jul 27.

Liu Z, Weaver J, De Papp A, Li Z, Martin J, Allen K, Hui S, Imel EA. Disparities in osteoporosis treatments. Osteoporosis International. 2016 Feb 1;27(2):509-19.

National Osteoporosis Foundation (NOF). What is Osteoporosis? http://nof.org/articles/7 (November 1, 2013)

Silverman, S.L., R.E. Madison. 1988. Decreased incidence of hip fracture in Hispanics, Asians, and blacks: California Hospital Discharge Data. Am J Public Health. 78:1482–83.

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

Measure Number (*if previously endorsed*)**:** 0053

Measure Title: Osteoporosis Management in Women Who Had a Fracture

Date of Submission: 4/9/2018

Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
2014 Submission:
N/A

1.3. What are the dates of the data used in testing? Click here to enter date range
2014 Submission:
Sample 1: January 1 to December 31, 2012.
Sample 2: July 2000 through December 2001

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

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

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
2014 Submission:
Sample 1: This measure was tested for reliability and meaningful difference in performance at the plan level using data from all Medicare health plans submitting HEDIS data for measurement year 2012. The plans were nationally representative and included 235 HMO plans and 112 PPO plans. The plans varied in size from a minimum of 30 eligible patients to over 6,441 within a single plan.
Sample 2: This measure was originally field tested in a sample of 5 health plans. The five plans were geographically diverse and included both HMOs and PPOs.

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*)
2014 Submission:
Sample 1: In 2012, HEDIS measures covered 8.7 million Medicare beneficiaries. Data is summarized at the health plan level for all Medicare plans submitting data for this measure for 2012. Patients included in the HEDIS data include a diverse representation of ages, race and diagnoses. The table below shows the average number of eligible patients per health plan and the standard deviation of that average across health plans.
Table 1: Sample 1 Average Eligible Population per Health Plan.

Product Type	Number of Plans	Average number of eligible patients per plan	Standard Deviation
Medicare	347	372	625

Sample 2: The sample from the field test conducted in five health plans included all women who experienced a fracture between July 2000 and June 2001. Table 2 below shows the number of women who experienced a fracture in each health plan by age.
Table 2: Sample 2 Eligible Population in each Field Test Health Plan

	Age	
Health Plan	67 and over	50 to 66

Plan A	202	207
Plan B	796	390
Plan C	613	201
Plan D	703	175
Plan E	876	199
Total	**3,190**	**1,172**

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.
2014 Submission:
Sample 1 was used to demonstrate reliability (beta-binomial calculation), construct validity (correlation analysis) and meaningful difference in performance.
Sample 2 was used to field test the measure, test item-level validity and exclusions.

Plan-level validity was also demonstrated through a systematic assessment of face validity. This measure was systematically evaluated for face validity with four panels of experts:

- The Osteoporosis Advisory Workgroup included 5 experts in geriatrics, endocrinology, and osteoporosis.
- The Geriatric MAP included 13 experts in geriatrics, including representation by consumers, health plans, health care providers and policy makers.
- The Technical Measurement Advisory Panel includes 14 members, including representation by health plans methodologists, clinicians and HEDIS auditors.
- NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 21 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

Per NQF instructions we have described the composition of the expert panels which assessed face validity for this measure. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panels.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
2018 Submission:
We did not analyze performance by social risk factors.

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*)
2014 Submission:

Plan-level reliability testing of performance measure score: In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient." This approach is also relevant to health plans and other accountable entities.

Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

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

2014 Submission:

Results of reliability testing of performance measure score: The table 3 below shows the results of the reliability testing of the performance measurement score in 2012.

Table 3: Reliability in Medicare Plans in 2012

# of plans	Overall Reliability Score	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile
347	.92	.81	.89	.95	.97	.99

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

2014 Submission:

Interpretation of measure score reliability testing: Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests this measure has very good reliability. The 10-90th percentile distribution of health plan level-reliability for this measure show nearly all health plans met or exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.9. Strong reliability is demonstrated with the majority of variance attributed to signal and not to noise.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

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

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source,*

relationship to another measure as expected; what statistical analysis was used)

2014 Submission:

Method of testing critical data element validity: To test the validity of plan administrative data for computing this measure, participating field test plans (Sample 2) selected a random sample of 100 patients from their administrative data file and reviewed their primary care physician medical records. This data was used as a gold standard to verify the completeness and accuracy of the administrative data concerning the date and type of fracture, clinical exclusions, and dates and types of treatment. Given the small sample size and high level of agreement (see 2b.2.3), no statistical test of agreement was performed.

Method of testing empirical validity: We tested for construct validity by exploring whether performance for this measure was correlated with a similar measure, Osteoporosis Testing in Older Women, in the most recent year of available HEDIS data (Sample 1). This measure assesses the proportion of women who report having ever received a bone mineral density test to check for osteoporosis. The measures focus on the same disorder, osteoporosis, in different populations. We hypothesized that these two measures would be positively correlated (i.e. plans that have high rates of performance for management of osteoporosis will also have high rates of performance for screening of osteoporosis.) To test this correlation we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

2018 Submission:

Method of assessing face validity: We describe below NCQA's process for both measure development, and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assures that measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2.

ICD-10 conversion: Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10:

All tools used for mapping/code identification from CMS ICD-10 website (http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html).

GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

Expert Participation:

The NCQA HEDIS Expert Coding Panel and NCQA's Diabetes Expert Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

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

2014 Submission:

Results critical data element validity test: The results in table 4 and 5 below show the number of numerator and denominator events identified in each field test plan's sample of 100 patients. The results demonstrate high agreement between medical records and administrative data.

Table 4: Numerator and Denominator events as identified by medical record and administrative data for adults age 65 and Older

	Denominator		Numerator		Rate	
Health Plan	MR	Admin	MR	Admin	MR	Admin
A	35	31	2	2	6%	6%
B	37	37	6	6	16%	16%
C	9	11	1	1	11%	9%
D	27	32	2	1	7%	3%
E	27	28	5	6	19%	21%
Total	135	139	16	16	12%	12%

Table 5: Numerator and Denominator events as identified by medical record and administrative data for adults age 50-64

	Denominator		Numerator		Rate	
Health Plan	MR	Admin	MR	Admin	MR	Admin
A	35	32	5	4	14%	13%
B	20	20	1	1	5%	5%
C	3	2	2	0	67%	--
D	1	1	0	0	--	--
E	8	6	2	2	25%	33%
Total	67	61	10	7	15%	11%

Results of empirical validity test:

The results in Table 6 indicated that for plan-level reporting this measure was significantly ($p<.05$) correlated with the Osteoporosis Testing measure (NQF #0037) in the direction that was hypothesized.

Table 6. Correlation between Osteoporosis Measures in Medicare Plans - 2012

Pearson Correlation Coefficient	
	Osteoporosis Testing in Older Women
Osteoporosis Management in Women who have had a Fracture	R=0.27305 (R Statistic) p<.0001 (significance)

Note: All correlations are significant at $p<.05$

Results of face validity assessment:

- Step 1: This measure was developed in 2002 to address under-diagnosis and treatment of osteoporosis in women who had fragility fractures. NCQA, along with the Osteoporosis Technical Subgroup and the Geriatric Measurement Advisory Panel, worked together to assess the most appropriate management steps for women who had a fragility fracture.
- Step 2: The measure was written and field-tested in 2002. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2003.
- Step 3: The measure was released for Public Comment in 2003 prior to publication in HEDIS. The CPM recommended moving this measure to first year data collection by a majority vote.
- Step 4: The measure was introduced in HEDIS 2004. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure to public reporting with a majority vote.

- Step 5: The measure was re-evaluated in 2013 and reviewed by the Osteoporosis Workgroup and the Geriatric Measurement Advisory Panel. The measure was presented to the CPM in January 2014 and proposed changes to the measure were posted for public comment February-March 2014. The CPM approved the proposed changes to the measure in May 2014 with a majority vote. These changes will go forward for use in HEDIS 2015.
- Conclusion: The measure was deemed to have the desirable attributes of a HEDIS measure in 2003 (relevance, scientific soundness, and feasibility).

Results of ICD-10 conversion:
Summary of Stakeholder Comments Received
NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. NCQA received comments from four organizations:

- Support recommendations.
- Questions about select codes.
- Recommended additional codes for consideration.

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?*)
2014 Submission:
Interpretation of data element validity testing: The results demonstrate near perfect agreement between medical records and administrative data. On average, health plans identified slightly more denominator events using administrative data and slightly more numerator events using medical record data. However, these differences were minor. We interpret this to suggest the administrative data elements used in this measure are valid compared to a gold standard medical record source.

Interpretation of empirical validity testing:
Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. *The results confirmed the hypothesis that this measure is correlated with the Osteoporosis Testing in Older Women (NQF #0037), suggesting they represent the* same underlying construct of quality of care for osteoporosis. Although the association was weak, it was significantly more than zero. A strong correlation would not be expected in this case due to the different denominators of these two measures.

2018 Submission:
Interpretation of face validity assessment:
NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Osteoporosis Management in Women Who Had a Fracture* is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)
2014 Submission:
At the time of field test, only one exclusion in the measure was tested, exclusion for women with prior treatment for osteoporosis (treatment on or within the 12 months prior to the fracture). The aim of testing exclusions in the field test data was to determine how common exclusions are in the eligible patient population and the impact of these exclusions

on denominator sizes and performance rates. Our results (detailed below) show differences in performance rates with and without exclusions and across data sources (administrative vs. medical record).

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*)
2014 Submission:
On average 34% of women 65+ who experienced a fracture met the exclusion criteria for current treatment (prescription for treatment in past 12 months); 51% of women 50-64 met the exclusion criteria. The application of the exclusion to the measure reduced rates by more than 60% for both age groups (see Table 7). At the time of the field test, Hormone Replacement Therapy (HRT) was considered a safe treatment for osteoporosis. More than half (64%) of women age 65+ who met the exclusion criteria were prescribed HRT. Almost all (92%) women 50-64 who met exclusion criteria were prescribed HRT.

Table 7: Exclusion for Treatment for Osteoporosis in prior 12 months

	Age 65+			Age 50-64		
	Number with current or prior RX	Rate with exclusion	Rate without exclusion	Number with current or prior RX	Rate with exclusion	Rate without exclusion
Plan A	115	11%	36%	234	11%	49%
Plan B	616	11%	41%	495	14%	53%
Plan C	328	12%	31%	162	13%	43%
Plan D	244	13%	31%	135	11%	41%
Plan E	383	14%	33%	178	18%	49%
Total	1686	13%	35%	1204	13%	48%

To determine the most appropriate data source for identifying exclusions, we compared medical records to administrative data in a sample of approximately 100 patients per plan. Across all three sites, the majority of exclusions could be identified through administrative data. Only 1% of records had an exclusion identified through medical record data alone (see Table 8).

Table 8: Exclusion for Treatment for Osteoporosis by Data Source

Plan Name	Patients	Admin Only	MR Only	Admin & MR	Neither
Plan A	116	6%	0%	40%	54%
Plan B	110	0%	0%	48%	52%
Plan C	100	82%	2%	5%	11%
Plan D	100	59%	0%	3%	38%
Plan E	100	62%	3%	4%	31%
Total	526	40%	1%	21%	38%

Admin: Administrative Data
MR: Medical Record

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*)
2014 Submission:
It is important to exclude women who have a prescription or therapy for osteoporosis treatment in the past 12 months so that the measure is focused on women who are not already on treatment at the time of the fracture. The exclusion looks back 12 months to identify a prescription for osteoporosis treatment because some osteoporosis treatments can be effective for up to 12 months. The field test identified that excluding women who have a prior prescription to treat osteoporosis in the past 12 months significantly impacts the measure, reducing rates by more than 60%. The test also

identified that administrative data was sufficient to identify this exclusion. Therefore, we determined this exclusion to be important and feasible.

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

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

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

☐ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

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.

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

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

1b)

2014 Submission:

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities.

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*)

2014 Submission:

Table 9: Variation in Performance across Health Plans in HEDIS (2012 data)

	Avg. EP	Mean Rate	SD	10th	25th	50th	75th	90th	IQR
Medicare Plans	372	23.1	13.7	12.2	15.0	19.1	25.9	40.5	10.9

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range

Table 10: T-test between two randomly selected health plans in HEDIS (2012 data)

	Plan Rate (25th Percentile)	Plan Rate (75th Percentile)	P-Value
Medicare Plans	12.1	36.8	.00003

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

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?*)

2014 Submission:

The results above indicate there is a 10.9% gap in performance between the 25th and 75th performing plans (see Table 9). The difference between the 25th and 75th percentile is statistically significant (see Table 10). This gap represents on average 40 more patients receiving bone mineral density testing or osteoporosis treatment following a fracture in high performing Medicare plans compared to low performing plans (estimated from average health plan eligible population).

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

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

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

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

used)
2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS
2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
2014 Submission:
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
2014 Submission:
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

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*)
2014 Submission:
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

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

Measure Number (*if previously endorsed*)**:** 0053

Measure Title: Osteoporosis Management in Women Who Had a Fracture

Date of Submission: 4/9/2018

Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

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

2014 Submission:

N/A

1.3. What are the dates of the data used in testing? Click here to enter date range

2014 Submission:

Sample 1: Testing of data element reliability was performed during field testing using 2009 medical record data.

Sample 2: Testing of performance variability was performed using 2012 performance data from the Physician Quality Reporting System.

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

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

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

2014 Submission:

Sample 1: This measure was tested for data element reliability at the provider level using field test data. To identify clinics for field testing, the American Academy of Orthopedic Surgeons (AAOS) posted an announcement online and also identified practices that were known through their previous work with the AAOS. Of the thirteen clinics who expressed an interest in the field-testing, two were chosen to participate. These two sites were chosen based on having participated in the 2009 Physician Quality Reporting Initiative (PQI) program with additional consideration given to balancing practice size, location, and use of an EHR or paper medical record. One site was located in New Mexico and one was located in South Carolina.

Sample 2: Reporting at the provider-level for this measure is collected through the Physician Quality Reporting System (PQRS). In 2012, 1730 providers nationwide reported on this measure. This reflects 0.8% of eligible providers.

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*)

2014 Submission:

Sample 1: Desired sample sizes for testing were calculated for this measure with 0.80 power, 0.05 significance, and testing for a kappa of substantial agreement (0.8) versus moderate agreement (0.4). Expected performance was conservatively assumed at 0.5, unless performance information was available for the measure. Based on these assumptions and calculations, the minimum number of patients needed for the sample was 38. A total of 39 patient records were reviewed across the two sites.

Sample 2: The number of beneficiaries reported on thought this measure for 2012 was 11,284.

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.

2014 Submission:

Sample 1 was used to test critical data element reliability.

Sample 2 was used to test meaningful differences in measure performance.

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

2018 Submission:

We did not analyze performance by social risk factors.

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*)

2014 Submission:

Critical data element reliability: Reliability was tested by assessing whether two abstractors, reviewing the same full medical (including both inpatient and outpatient notes), would come to the same conclusion as to the patient meeting the measure, not meeting the measure, or qualifying as an exception. Two abstractors independently assessed whether patients met numerator inclusion criteria for each case that met denominator inclusion criteria. Following the data abstraction, the mismatches were tallied. Agreement between abstractors was measured using the kappa statistic (a measure of agreement adjusted for agreement that can occur by chance).

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*)

2014 Submission:

Denominator: Agreement between the two independent reviewers was 100% for the denominator data element; with reviewers agreeing 36/39 cases met the denominator criteria. The Kappa was 1.00, which indicates that there was perfect agreement that the two abstractors came to the same conclusion as to patients who met the denominator.

Exceptions: Agreement between the two independent reviewers was 96.2% for the exception data element. The reviewers disagreed about 1/36 cases where one reviewer found evidence of the exception criteria and one reviewer found no evidence of the exception criteria. The Kappa was .65 for this data element. The reviewers met to reconcile their differences and determined the exception criteria was met in 1 case.

Numerator: Agreement between the two reviewers was 83.3% with agreement that the numerator criteria was met in 4/36 cases and not met in 24/36 cases. The reviewers disagreed about 6/36 cases where one reviewer found evidence that the numerator criteria was met and one review did not find evidence in the medical record that numerator criteria was met. A Kappa statistic was calculated to demonstrate the degree of agreement adjusted for chance (K=0.47; 95% CI: 0.11-.83). The two reviewers met to reconcile their differences and determined 5/36 cases met the numerator criteria for a performance rate of 13.9%.

Table 1 below displays the overall inter-rater reliability for the all the measure components combined. Concordance between the abstractors is 82% with moderate agreement above what would be expected (κ 95% confidence interval 0.35-.88).

Table 1: Inter-rater reliability of measure components

		Reviewer B				
		Not Study Eligible	Not Met	Met	Exclusion	Total
Reviewer A	Not Study Eligible	3	0	0	0	3
	Not Met	0	24	2	1	27
	Met	0	4	4	0	8
	Exclusion	0	0	0	1	1
	Total	3	28	6	2	39

"Not Study Eligible" means that the denominator criteria were not met.

"Not Met" means denominator criteria were met, numerator criteria were not met and exceptions (exclusions) did not apply.

"Met" means denominator criteria were met and numerator criteria were met.

"Exclusion" means denominator criteria were met, numerator criteria were not met and exclusion applied.

Kappa Coefficient	0.61

Kappa LL (95% Confidence Interval)	0.35
Kappa UL (95% Confidence Interval)	0.88
Observed Agreement Rate	0.82
Expected Agreement Rate	0.54

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?*)

2014 Submission:

Interpretation of data element reliability testing: The below scale was used in the field test to interpret the kappa score. The denominator had a kappa score of 1.00, which indicates that there was perfect agreement that the two abstractors came to the same conclusion as to patients who met the denominator. The numerator had a kappa score of .47, which indicates that there was moderate agreement that the two abstractors came to the same conclusion as to patients who met the numerator. The exceptions part of this measure had a kappa score of .65, which indicates that there was substantial agreement that the two abstractors came to the same conclusion as to patients who met the exception criteria. Across all data elements the kappa was 0.61 indicating substantial agreement between raters. This suggests the measure elements can be reliably abstracted from medical records.

Kappa	Strength of Agreement
0.00	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect

Landis, J.R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159—174

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

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

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

2018 Submission:

There are no updates to the validity testing for this measure since the last submission. The only publicly available data for this measure are from reporting in the CMS Quality Payment Program, however these data are not constructed in a way that allows NCQA to test empirical validity of the measure.

2014 Submission:

Critical Data Element Validity: The testing conducted for this measure by the AMA/Physician Consortium for Performance Improvement (PCPI) is described above under "Reliability." This testing demonstrates inter-rater reliability of two reviewers using the same measure specification to draw conclusions from the same "gold-standard" data source (e.g. medical record). Reliability testing demonstrated that two independent reviewers looking at the same full medical record had high agreement on every data element and the overall performance measure score. We believe this testing

demonstrates not only data element reliability but also validity, that is to say the accuracy of the measure specification to identify all data elements from the medical record.

Assessment of face validity: The AMA-convened Physician Consortium for Performance Improvement (PCPI) oversees the measure development process of clinically relevant physician-level performance measures. To assess the face validity of measures, PCPI follows a standardized process for measure development which includes:

- Convening cross-specialty, multidisciplinary work groups to assess the face and content validity of each measure. The groups establish the measure's ability to capture what it is designed to capture using a consensus process that consists of input from multiple stakeholders, including practicing physicians and experts with technical measure expertise.
- Review of the evidence, gaps in care and potential for impact of the measure:
    - Consider existing guideline recommendations and the strength of evidence
    - Consider gaps in care, variation, cost and frequency data
- Posting the draft measure for a 30-day public comment period. The PCPI solicits feedback from PCPI members, quality improvement collaboratives, providers, consumers, public/private purchasers and others with an interest in the measure.
- The PCPI work group reviews comments received, revises and modifies the draft performance measures as deemed appropriate by the work group. The public comments and responses are posted to the PCPI website as part of the voting process.
- Final vote by PCPI members eligible to vote. The PCPI encourages all voting member organizations to vote so the required quorum is met.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
2014 Submission:
Results of face validity assessment: This measure was reviewed and developed by a joint work group that included experts in osteoporosis treatment as well as representatives from the following organizations: American Academy of Family Physicians; American Academy of Orthopaedic Surgeons; American Association of Clinical Endocrinologists; American College of Rheumatology; The Endocrine Society; American Medical Association; National Osteoporosis Foundation; National Committee for Quality Assurance; and The Joint Commission. The joint work group members came to consensus on the final recommended specification for this measure in October 2006. **See section Ad. 1. Workgroup/Expert Panel Involved in Measure Development** for a list of participants of the Osteoporosis Work Group.

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?*)
2014 Submission:
Interpretation of face validity assessment: These results indicate that the multiple experts and stakeholders concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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*)
2018 Submission:
The exclusions for this measure are based on clearly specified codes that indicate the patient received hospice services or resided long-term in an institutional setting during the measurement period, patients who received previous pharmacologic therapy to treat osteoporosis in the previous 12 months or patients who had a bone mineral density test in the two years prior to the fracture. While these codes have not been specifically tested in the context of this measure,

they are considered valid for identifying patients who should be excluded from the measure. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons.

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*)

2018 Submission:

NA

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*)

2018 Submission:

NA

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

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

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

☐ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

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.

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

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

_____

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

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

2014 Submission:

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

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*)

2014 Submission:

Table 4: Variation in Performance across Providers (2012 data)

Mean Rate	EP	10th	25th	50th	75th	90th	IQR
70.0	11,284	0.0	25.0	100.0	100.0	100.0	75.0

EP: Number of patients meeting denominator criteria across all providers submitting data to the Physician Quality Reporting System on this measure

IQR: Interquartile range

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?*)

2014 Submission:

The results above indicate there is a large gap in performance between providers at the 25th and 75th percentiles. This demonstrates a large variation in performance and significant room for improvement on this measure for many providers. It should be noted that performance data from the PQRS program does not reflect performance system wide because physicians have the option to report. We look forward to more detailed performance reports from PQRS that may demonstrate longitudinal provider-specific performance improvements.

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

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

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

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

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

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

for the test conducted)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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*)

2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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*)

2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

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):*
Endocrine, Musculoskeletal, Musculoskeletal : Falls and Traumatic Injury, Musculoskeletal : Osteoporosis

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

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

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.)*
N/A

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0053_OMW_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

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

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Added an exclusion for patients 65 years of age and older living long-term in institutional settings.
Added an exclusion for patients receiving hospice care.
There would be no benefit to assessing older women in hospice care to see whether they had a bone mineral density test to screen for osteoporosis. Additionall, getting a bone mineral density test to check for osteoporosis typically requires transportation to a health care facility, which may be burdensome for older adults living long-term in institutional settings who may also have trouble tolerating the medications used to treat osteoporosis.

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

rationale for the measure.
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who received either a bone mineral density test or a prescription for a drug to treat osteoporosis after a fracture occurs.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who received either a bone mineral density test or a prescription for a drug to treat osteoporosis in the six months after a fracture. Appropriate testing or treatment for osteoporosis after the fracture is defined by any of the following criteria:
- A bone mineral density test (see Table OMW-X) in any setting, on earliest date of service with the diagnosis of fracture or in the 180-day (6-month) period after the fracture. If the earliest date of service with the diagnosis of fracture was during an inpatient stay, a bone mineral density test taking place during the inpatient stay counts.
- Osteoporosis therapy, including long-acting injectables, on the earliest date of service with the diagnosis of fracture or in the 180-day (6-month) period after the fracture. If the earliest date of service with the diagnosis of fracture was an inpatient stay, long-acting osteoporosis medication received during the inpatient stay counts.
- A dispensed prescription to treat osteoporosis (see Table OMW-C) on the earliest date of service with the diagnosis of fracture or in the 180-day (6-month) period after the fracture.

Table OMW-X: Bone Mineral Density Tests
Central dual-energy x-ray absorptiometry, computed tomography, single energy x-ray absorptiometry, ultrasound

Table OMW-C: Osteoporosis Medication
Biphosphates: Alendronate, Alendronate-cholecalciferol, Ibandronate, Risedronate, Zoledronic acid
Other: Calcitonin, Denosumab, Raloxifene, Teriparatide

The numerator for this measure can be identified using either administrative claims or review of medical records. The following criteria are used to identify the numerator criteria for each method. *Note this measure has been tested using medical record review at the physician level and administrative data at the health plan level.

For Medical Record Review Methodology (Physician Level)
When using the medical record as the data source, the numerator criteria is met by documentation that a Bone Mineral Density Test was performed or an osteoporosis therapy was prescribed. This may include a prescription given to patient for treatment of osteoporosis at one or more encounters during the reporting period. This measure is also collected in the Quality Payment Program, previously referred to as the Physician Quality Reporting System, using G-codes specific to the quality measure:
- 3095F Central Dual-energy X-Ray Absorptiometry (DXA) results documented
- G8633 Pharmacologic therapy (other than minerals/vitamins) for osteoporosis prescribed

For Administrative Methodology (Health Plan Level)
When using administrative claims as the data source, the numerator criteria is met by one or more codes in the following value sets:
Bone Mineral Density Tests Value Set
Osteoporosis Medications Value Set
A pharmacy claim for a medication listed in Table OMW-C
See S.2b. (Data Dictionary Code Table) for all value sets.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women who experienced a fracture, except fractures of the finger, toe, face or skull. Three denominator age strata are reported for this measure:
Women age 50-64
Women age 65-85
Women age 50-85

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The denominator for this measure is identified by administrative codes which are specific to the level of reporting. When reporting this measure at the health plan level include all individuals with fractures enrolled in the health plan (i.e. all individuals with encounters for fractures in the health plan – inpatient and outpatient). When reporting this measure at the physician level include all individuals with fractures seen by the eligible provider (i.e., all individuals with encounters for fracture with the eligible provider).

Health Plan Level Denominator Details:
Women who had an outpatient visit (see Outpatient Value Set), an observation visit (see Observation Value Set), an ED visit (see ED Value Set), a nonacute inpatient encounter (see Nonacute Inpatient Value Set) or an acute inpatient encounter (see Acute Inpatient Value Set) for a fracture (see Fractures Value Set) during the 12-month window that begins on July 1 of the year prior to the measurement year and ends on June 30 of the measurement year. This is the index fracture. If the patient had more than one fracture during the intake period, include only the first fracture. See S.2b. (Data Dictionary Code Table) for all value sets.

Physician Level Denominator Details:
Women who had a documented patient encounter (See Table 1 for encounter codes) with a fracture diagnosis (See Fracture Value Set).

Table 1: Patient encounter during the reporting period:
CPT Service codes: 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, G0402
CPT Procedure codes: 22310, 22315, 22318, 22319, 22325, 22326, 22327, 22510, 22511, 22513, 22514, 25600, 25605, 25606, 25607, 25608, 25609, 27230, 27232, 27235, 27236, 27238, 27240, 27244, 27245, 27246, 27248

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
- Exclude women who had a bone mineral density test during the 24 months prior to the index fracture.
- Exclude women who had a claim/encounter for osteoporosis treatment during 12 months prior to the index fracture.
- Exclude women who received a dispensed prescription or had an active prescription to treat osteoporosis during the 12 months prior to the index fracture.
- Exclude women who are enrolled in a Medicare Institutional Special Needs Plan (I-SNP) or living long-term in an institution any time during the measurement year.
- Exclude women receiving hospice care during the measurement year.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
1) Exclude patients with a previous fracture: patients with an outpatient visit (see Outpatient Value Set), an observation visit (see Observation Value Set), an ED visit (see ED Value Set), a nonacute inpatient encounter (see Nonacute Inpatient Value Set) or an acute inpatient encounter (see Acute Inpatient Value Set) for a fracture (see Fractures Value Set) during the 60 days (2 months) prior to the earliest date of service with a diagnosis of fracture. For index fractures requiring an inpatient stay, use the admission date as the earliest date of service with a diagnosis of fracture. For direct transfers, use the first admission date as the earliest date of service with a diagnosis of fracture.
2) Exclude patients who had a Bone Mineral Density test (see Bone Mineral Density Tests Value Set) during the 730 days (24 months) prior to the earliest date of service with a diagnosis of fracture.
3) Exclude patients who had a claim/encounter for osteoporosis therapy (see Osteoporosis Medications Value Set) or received a dispensed prescription to treat osteoporosis (see Table OMW-C) during the 365 days (12 months) prior to the earliest date of service with a diagnosis of fracture.
4) Exclude patients who live long-term in Institutional settings (as identified by the LTI flag in the Medicare Part C monthly membership file) or are enrolled in a Medicare Institutional Special Needs Plan during the measurement year.
5) Exclude patients who are in hospice care during the measurement year (as identified by the Medicare plan's enrollment file).

Table OMW-C: Osteoporosis Therapies
Alendronate, Alendronate-cholecalciferol, Ibandronate, Risedronate, Zoledronic acid, Calcitonin, Denosumab, Raloxifene, Teriparatide

The denominator exclusions for this measure can be identified using administrative claims, health plan enrollment data or review of medical record. The following criteria are used to identify the denominator exclusion criteria for each method. *Note this measure has been tested using medical record review at the physician level and administrative data at the health plan level.

For Medical Record Review Methodology (Physician Level)
When using the medical record as the data source, the denominator exclusion criteria can be met by documentation that a previous fracture occurred, a bone mineral density test was performed or an osteoporosis therapy was prescribed during the specified timeframe prior to the fracture. In the Physician Quality Reporting System (PQRS) this exclusion is collected using G-codes specific to quality measurement:
- 3095F or 4005F with 1P: Documentation of medical reason(s) for not performing a bone mineral density test or not prescribing pharmacologic therapy for osteoporosis (i.e. history of fracture in 60 days prior to index fracture, bone mineral density test in 24 months prior to index fracture, or pharmacologic treatment for osteoporosis in 12 months prior to index fracture).

For Administrative Methodology (Health Plan Level)
When using administrative claims as the data source, the denominator exclusion criteria is met using the following value sets referenced above during the specified time frame prior to the fracture.
Outpatient Value Set
ED Value Set
Nonacute Inpatient Value Set
Acute Inpatient Value Set
Fractures Value Set
Bone Mineral Density Tests Value Set
Osteoporosis Medications Value Set
See S.2b. (Data Dictionary Code Table) for all 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.)*
N/A

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 = Higher 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.)*
Health Plan Level:
Step 1: Identify all female patients who had a new fracture during the intake period (12-month window that begins on July 1 of the year prior to the measurement year and ends on June 30 of the measurement year).

Step 2: Exclude patients who had previous bone mineral density test and patients who had previous osteoporosis treatment. Also exclude patients living long-term in institutional settings and patients receiving hospice care.

Step 3: Of those patients remaining after Step 2 (i.e., the denominator), identify those who received bone mineral density testing or osteoporosis treatment in the 6-month period following the fracture.

Step 4: To calculate the rate, take the number of patients who received testing or treatment and divide by the number of people calculated to be in the denominator.

Physician Level:
Step 1: Identify all female patients in each age strata who had a documented patient encounter with the eligible provider with a new diagnosis of fracture.
Step 2: Exclude patients who had who had previous bone mineral density test and patients who had previous osteoporosis treatment. Also exclude patients living long-term in institutional settings and patients receiving hospice care.

Step 3: Of those patients remaining after Step 2 (i.e., the denominator), identify all patients who had a documented bone mineral density test or pharmacologic treatment after the fracture.

Step 4: To calculate the rate, take the number of patients who received testing or pharmacologic treatment and divide by the number of people calculated to be in the denominator.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Electronic Health Records, Paper Medical 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.
Health Plan Level:
This measure is based on administrative claims collected in the course of providing care to health plan patients. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Maintenance Organizations and Preferred Provider Organizations via NCQA's online data submission system.

Physician Level:
This measure is based on administrative claims to identify the eligible population and medical record documentation collected in the course of providing care to health plan patients to identify the numerator. In the Quality Payment Program, this measure is collected using G-codes specific to quality measurement.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Health Plan, Integrated Delivery System

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.)*
N/A

2. Validity – See attached Measure Testing Submission Form
0053_-_Testing_Form_v7.1_FINAL-636596510946647046.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.

Yes

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.

Yes

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.

No - This measure is not risk-adjusted

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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**.

Some data elements are in defined fields in electronic sources

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).

To allow for widespread reporting across physicians and clinical practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records).

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

Attachment:

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Health Plan Level:
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.
The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

Physician Level:
Feedback on use of this measure in CMS PQRS program has been positive with few questions raised by participating clinicians to the CMS vendor. NCQA works with the CMS vendor to review any questions or issues raised with the measure on a bi-weekly basis.

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)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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 Health Plan Rantings http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality CMS Quality Payment Program https://qpp.cms.gov Health Plan Rantings http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality CMS Quality Payment Program https://qpp.cms.gov Payment Program Medicare STARS https://www.medicare.gov/find-a-plan/questions/home.aspx CMS Quality Payment Program https://qpp.cms.gov Medicare STARS https://www.medicare.gov/find-a-plan/questions/home.aspx CMS Quality Payment Program https://qpp.cms.gov Regulatory and Accreditation Programs NCQA Accreditation http://www.ncqa.org/tabid/123/Default.aspx HEDIS ACO http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx NCQA Accreditation http://www.ncqa.org/tabid/123/Default.aspx HEDIS ACO http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Quality Improvement (external benchmarking to organizations) CMS Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass Annual State of Health Care Quality

	http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality

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

HEALTH PLAN LEVEL USE:
NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population.

CMS MEDICARE ADVANTAGE STAR RATING: This measure is included in the composite Medicare Advantage Star Rating. CMS calculates a Star Rating (1-5) for all Medicare Advantage health plans based on 53 performance measures. Medicare beneficiaries can view the star rating and individual measure scores on the CMS Plan Compare website. The Star Rating is also used to calculate bonus payments to health plans with excellent performance. The Medicare Advantage Plan Rating program covers 11.5 million Medicare beneficiaries in 455 health plans across all 50 states.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported by WebMD and on the NCQA website. These rantings are based on performance on HEDIS measures among other factors. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.

NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole person care coordination throughout the health care system.

NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries. Health plans are scored based on performance compared to benchmarks.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

PHYSICIAN LEVEL USE
CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

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?*)
N/A

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.*)
N/A

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs.

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.**
During the measure's last major update in 2014, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure.

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.
Health Plan Level: From 2014 to 2016, the average performance rate has increased by four percentage points. Since 2013, rates have increased about 18.4 percent for health plans in the 90th percentile (see section 1b.2 for summary of data from health plans). In 2016, a total of 277 Medicare health plans reported data on this measure. These data are nationally representative.

Physician Level: From 2009-2012 the average performance rate has increased by 13.5 percent, which shows steady improvement amongst those providers who chose to report on this measure. In 2012, there were 204, 369 eligible providers who were able to report on this measure and only 0.8% choose to report. Therefore, the 2012 average performance rate is reflective of less than one percent of Medicare providers.

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.**
There is a possibility that this measure may inadvertently increase the overuse of bone mineral density tests and approved treatments for osteoporosis and fractures, especially in those who have a limited life expectancy. Although the population of women with recent osteoporotic fractures is least likely to be associated with overuse, the asymptomatic population is more prone to this. To help minimize this, we have an upper age limit of 85 for this measure and specific exclusions for those in hospice care and those living long-term in institutional settings. NCQA is also currently exploring additional exclusions to remove patients with advanced illness from this measure. These exclusions focus the measure on the population that is most likely to benefit from screening and treatment.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits for this measure during implementation.

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)**
0037 : Osteoporosis Testing in Older Women (OTO)
0046 : Screening for Osteoporosis for Women 65-85 Years of Age
2416 : Laboratory Investigation for Secondary Causes of Fracture
2417 : Risk Assessment/Treatment After Fracture

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

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.**
Insufficient Space - please see 5b.1.

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.)
Response to 5a.2 (insufficient space above): There are multiple measures of osteoporosis prevention and management. During the last measure update in 2014, this measure was harmonized to align with applicable existing NQF-endorsed osteoporosis measures where possible given the different measure focus, methods of data collection and level of accountability. Below we describe the harmonization between this measure (0053) and the most closely related measures, 0037, 0046, 2416, 2417.

NCQA OWNED RELATED MEASURES
0037: Osteoporosis Testing in Older Women
0046: Screening for Osteoporosis for Women 65-85 Years of Age
Measures 0037 and 0046 assess the number of women 65-85 who report ever having received a bone density test to check for osteoporosis. These measures focus on screening for osteoporosis in the general population, whereas measure 0053 is focused on secondary prevention in a population of women who have experienced a fracture. Therefore, we consider these measures to be related but not competing. The differences between these two measures are reflective of the different guidelines for general population screening and secondary prevention. Where it is appropriate to the measure focus and evidence, we have aligned the measures.

OTHER RELATED MEASURES
The other osteoporosis management related measures are more narrowly focused than the NCQA measures. These measures (2416, 2417) are hospital-level accountability measures and focus solely on women who were hospitalized for fractures.

2416: Laboratory Investigation for Secondary Causes of Fracture
Measure 2416 assesses the percentage of patients age 50 and over who were hospitalized for a fragility fracture and had the appropriate laboratory investigation for secondary causes of fracture ordered or performed prior to discharge from an inpatient hospitalization. This measure has a different focus from measure 0053 (identifying cause of fracture as opposed to screening/treatment for osteoporosis). While the target population of this measure overlaps with the target population of 0053, measure 2416 is restricted to fractures that require hospitalization whereas 0053 focuses on a broader population. Therefore, we consider these measures to be related but not competing. Measure 2416 captures some of the same quality focus as 0053 but is designed to be appropriate for hospital-level accountability and is therefore restricted to hospitalized individuals. The differences between this measure and 0053 are reflective of the different measure intents and level of accountability.

2417: Risk Assessment/Treatment After Fracture
Measure 2417 assesses the number of patients age 50 and over who were hospitalized for a fragility fracture and have either a dual-energy x-ray absorptiometry (DXA) scan ordered or performed, a prescription for FDA-approved pharmacotherapy, or are linked to a fracture liaison service prior to discharge from an inpatient hospitalization. If DXA is not available and documented, then any other specified fracture risk assessment method may be ordered or performed. This measure has a similar focus to 0053 and an overlapping target population (individuals hospitalized for a fragility fracture). Therefore, this measure could be considered competing with 0053; however, 2417 is designed to focus on hospital-level accountability and therefore is only inclusive of populations and services provided within the hospital setting. Measure 0053 is designed to be broader and capture both outpatient and inpatient populations and services.

Response to 5b.1: This measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure:2417 Risk Assessment/Treatment After Fracture.

Measure 0053 is designed to be as broad as possible to include the largest possible population (all women age 50 and over with a fracture other than face, finger, toe, and skull) and include the broadest possible settings of care (inpatient and outpatient). The measure is designed for both health plan and outpatient physician level accountability. It is focused on guideline recommended care for osteoporosis management after a fracture. Measure 2417 is designed to be appropriate for hospital-level accountability and therefore focuses on a smaller population (all patients 50 and over hospitalized for a fragility fracture) and includes a single setting of care (inpatient). While some post-fracture care occurs in the inpatient setting, much of the responsibility for providing follow-up care for osteoporosis management in women rests with the outpatient care system and providers. Additionally, many patients who suffer a fracture may not be treated with an inpatient hospitalization. Therefore, it is important to have a measure that captures a broader population and settings of care for osteoporosis management following a fracture.

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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
Geriatric Measurement Advisory Panel (GMAP)
Wade Aubry, MD, University of California, San Francisco
Arlene S Bierman, MD, MS, Agency for Healthcare Research and Quality (AHRQ)
Patricia A. Bomba, MD, MACP, Excellus BlueCross BlueShield
Nicole Brandt, PharmD, MBA, BCGP, BCPP, FASCP, University of Maryland, School of Pharmacy
Jennie Chin Hansen, RN, Geriatric Expert
Joyce Dubow, MUP, Consumer Representative
Gustavo Ferrer, MD, Aventura Hospital
Peter Hollmann, MD, University Medicine
Jeffrey Kelman, MD, MMSc, Centers for Medicare & Medicaid Services (CMS)
Karen Nichols, MD, AmeriHealth Caritas Family of Companies Steven Phillips, MD, CMD, Geriatric Specialty Care
Jane Sung, JD, AARP
Eric G Tangalos, MD, FACP, AGSF, CMD, Mayo Clinic
Dirk Wales, MD, PsyD, Cigna HealthSpring
Joan Weiss, PhD, RN, CRNP, Health Resources and Services Administration
Neil Wenger, MD, UCLA Division of General Internal Medicine and RAND

Osteoporosis Advisory Workgroup
Joyce Dubow, MUP, Consumer Representative
Margery Gass, MD, NCMP, The North American Menopause Society
Peter Hollmann, MD, University Medicine
Steven Petak, MD, MACE, JD, Endocrinologist, Houston Methodist Hospital Academic Associates
Kenneth G. Saag, MD, MSc, Divison of Clinical Immunology and Rheumatology, University of Alabama at Birmingham

Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
Jonathan Darer, MD, MPH, Medicalis
Helen Darling, MA, City of Washington, DC
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare & Medicaid Services
David Grossman, MD, MPH, Kaiser Permanente Washington
Christine S. Hunter, MD, US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Nancy Lane, PhD, Newton, MA
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System

Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric Schneider, MD, MSc, FACP, The Commonwealth Fund
Marcus Thygeson, MD, MPH, San Rafael, CA
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms
Lina Walker, PhD, AARP

Physician Consortium for Performance Improvement Osteoporosis Workgroup
Steven Petak, MD, MACE, JD, Endocrinologist, Houston Methodist Hospital Academic Associates
Kenneth G. Saag, MD, MSc, Divison of Clinical Immunology and Rheumatology, University of Alabama at Birmingham
Robert Alder, MD
H. Chris Alexander, III, MD, FACP
Donald Bachman, MD, FACR
Joel Brill, MD
Jan Busby-Whitehead, MD
Thomas Dent, MD
Nancy Dolan, MD
Leonie Gordon, MB, ChB
Thomas Griebling, MD
Richard Hellman, MD, FACP, FACE
Marc C. Hochberg, MD, MPH
C. Conrad Johnston, Jr. MD
Joseph Lane, MD
Leon Lenchik, MD
Bonnie McCafferty, MD, MSPH
Michael Maricic, MD
Michael L. O'Dell, MD, MSHA, FAAFO
Sam J.W. Romeo, MD, MBA
Frank Salvi, MD, MS
Joseph Shaker, MD
Madhavi Vemireddy, MD
David Wong, MD, MSc, FRS(C)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2003

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if clinical guidelines or evidence has changed significantly

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.
Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure. These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0055
Corresponding Measures:
Measure Title: Comprehensive Diabetes Care: Eye Exam (retinal) performed
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who had an eye exam (retinal) performed.
Developer Rationale: This measure promotes regular eye examinations in diabetic adults (ages 18-75). Diabetic retinopathy and vision loss are complications from diabetes. Adults with diabetes that do not receive regular retinal examinations are at a higher risk for developing these vision complications. Vision screenings are part of high quality care for patients with diabetes.

Numerator Statement: Patients who received an eye screening for diabetic retinal disease. This includes people with diabetes who had the following:
-a retinal or dilated eye exam by an eye care professional (optometrists or ophthalmologist) in the measurement year
–a negative retinal exam or dilated eye exam (negative for retinopathy) by an eye care professional in the year prior to the measurement year.
-Bilateral eye enucleation anytime during the patient's history through December 31 of the measurement year

For exams performed in the year prior to the measurement year, a result must be available.
Denominator Statement: Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Exclude patients who did not have a diagnosis of diabetes, in any setting, AND who had a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year
-Exclude patients 65 and older with an advanced illness condition and frailty

Measure Type: Process
Data Source: Claims, Electronic Health Data, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 or **Summary of prior review in [year]**

The developer provided updated guidelines from the American Diabetes Association (ADA) (2018) including recommendations for the following:

- Adults with type 1 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist within 5 years after the onset of diabetes. **B**
- Patients with type 2 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist at the time of the diabetes diagnosis. **B**
- If there is no evidence of retinopathy for one or more annual eye exam and glycemia is well controlled, then exams every 1–2 years may be considered. If any level of diabetic retinopathy is present, subsequent dilated retinal examinations should be repeated at least annually by an ophthalmologist or optometrist. If retinopathy is progressing or sight threatening, then examinations will be required more frequently. **B**
- While retinal photography may serve as a screening tool for retinopathy, it is not a substitute for a comprehensive eye exam. **E**
- Women with preexisting type 1 or type 2 diabetes who are planning pregnancy or who are pregnant should be counseled on the risk of development and/or progression of diabetic retinopathy. **B**
- Eye examination should occur before pregnancy in the first trimester in patients with preexisting type 1 or type 2 diabetes, and then patients should be monitored every trimester and for 1 year postpartum as indicated by the degree of retinopathy. **B**
- Level of evidence and description:
    - B:
        - Supportive evidence from well-conducted cohort studies, including:
        - Evidence from a well-conducted prospective cohort study or registry
        - Evidence from a well-conducted meta-analysis of cohort studies
        - Supportive evidence from a well-conducted case-control study
    - E:
        - Expert consensus or clinical experience

The developer provided updated guidelines from the American Academy of Ophthalmology (AAO) (2017) including recommendations for the following:

- Table 3 Recommended Eye Examination for Patients with Diabetes Mellitus and No Diabetic Retinopathy

Diabetes Type	Recommended Initial Evaluation	Recommended Follow up*
Type 1	5 years after diagnosis (II++;Good, Strong)	Yearly (III; Good; Strong)
Type 2	At time of diagnosis (II+; Good Strong)	Yearly (III; Good; Strong)
Pregnancy (type 1 or type 2)	Soon after conception and early in the first trimester (III; Good; Strong)	No retinopathy to mild or moderate NPDR: every 3-12 months (III, Good, Strong) Severe NPDR or worse: every 1-3 months (III, Good, Strong)

- To rate individual studies, a scale based on Scottish Intercollegiate Guideline Network (SIGN) is used. The definition and levels of evidence to rate individual studies are as follows:
    - III Nonanalytic studies (e.g., case reports, case series)
- Recommendations for care are formed based on the body of the evidence. The body of evidence quality ratings are defined by GRADE as follows:
    - Good quality: Further research is very unlikely to change our confidence in the estimate of effect
- Key Recommendations for care are defined by GRADE as follows:
    - Strong recommendations: Used when the desirable effects of an intervention clearly outweigh the undesirable effects or clearly do not

The developer provided updated guidelines from the American Geriatrics Society (AGS) (2013) including recommendations for the following:

- "1. Older adults with new-onset DM should have an initial screening dilated-eye examination with funduscopy performed by an eye care specialist." **(Level I, Grade B)**
- "2. Older adults with DM and who are at high risk for eye disease (symptoms of eye disease present; evidence of retinopathy, glaucoma, or cataracts on an initial dilated-eye examination or subsequent examinations during the prior 2 years; A1C $\geq$ 8.0%; type 1 DM; or blood pressure $\geq$ 140/80) on the prior examination should have a screening dilated-eye examination performed by an eye-care specialist with funduscopy training at least annually. Persons at lower risk or after one or more normal eye examinations may have a dilated-eye examination at least every 2 years." **(Level II, Grade B)**
- Quality of Evidence
    - Level I: Evidence from at least one properly randomized controlled trial
    - Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytical studies, from multiple time-series, or from dramatic results in uncontrolled experiments
- Strength of Evidence
    - B: Moderate evidence to support the use of a recommendation clinicians "should do this most of the time"

The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the ADA and AGS guidelines, as this information was not available. In lieu of this, the developer cited two systematic reviews to support the recommendations (AACE Diabetes Care Plan Guidelines, 2011; Zhang et al., 2010)

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided updated guidelines from 2013 submission and cited two systematic reviews to support the guidelines recommendations.

Exception to evidence

NA

Questions for the Committee:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
    - *What is the relationship of this measure to patient outcomes?*
    - *How strong is the evidence for this relationship?*

- *Is the evidence directly applicable to the process of care being measured?*
- *If derived from patient report, does the target population value the measured process or structure and find it meaningful?*

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

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

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

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

- Developer provided performance data also from the 2015 PQRS reporting year with a mean of 78.1% and a standard deviation of 28.3%.
- Developer provided performance data for the NCQA's Diabetes Recognition Program (DRP) from 2015, 2016, and 2017.
  - Mean: 61.4% (2014) to 62.8% (2016)
  - Standard deviation: 24.3% (2014) to 21.3% (2016)
- Developer provided performance data extracted from HEDIS data, stratified by commercial health plan, Medicare, and Medicaid from 2014, 2015, and 2016.
  - Commercial performance
    - Mean: 52.6% (2014) to 50.5% (2016)
    - Standard Deviation: 12.3% (2014) to 12.6% (2016)
  - Medicare performance
    - Mean: 68.5% (2014) to 70.2% (2016)
    - Standard Deviation: 11.5% (2014) to 11.0% (2016)
  - Medicaid performance
    - Mean: 54.4% (2014) to 54.9% (2016)
    - Standard Deviation: 11.6% (2014) to 11.7% (2016)

Disparities
The developer did not provide disparities data for the measure. The developer cited two cross-sectional studies examining data from the Medical Expenditure Panel Survey.

- The 2014 study, using data from 2002-2009 found that that racial disparities and other influencing factors have an impact on rates of eye examinations among patients with diabetes and there needs to be more efforts to improve screening and testing of diabetic retinopathy among minorities (Shi et al).
- The 2018 study, using data from 2013, noted that improvement in quality in diabetes care will help reduce diabetes complications and mortality (Canedo et al.)

Questions for the Committee:
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The evidence is stronger than previous reviews. There is no need to repeat the discussion and to vote on the evidence.
- The developer provided updated guidelines from 2013 submission and cited two systematic reviews to support the guidelines recommendations. American Academy of Ophthalmology (AAO) (2017) and the 2018 American Diabetes Association guidelines. The American Geriatrics Society (AGS) (2013) guidelines were also provided.
- Dilated retinal exam informs early diagnosis and treatment of retinopathy in diabetic individuals: vision can be preserved or progression can be slowed. The measure applies directly to diabetic retinopathy The evidence base is updated, strengthened and is moderately strong.
- The evidence provided seems to target diagnosis and then less frequent when results are normal and glycemia is under control – what is the need for the annual exam then for patients who are controlled with a history of normal results? Is there evidence for an annual exam for controlled patients? Adding another visit could mean another day off of work for patients.
- Developer provided updated evidence that is directionally the same and complete
- Evidence supports rationale (new guidelines from American Diabetes Association and American Academy of Ophthalmology and American Geriatrics Society, 2 systematic reviews so support guideline recommendations)
- Relationship of measure to outcomes: see rationale
- Strong evidence: moderate (grade B or II)
- Evidence applicable to process of care being measured? Yes
- Therefore, no need for repeat discussion and vote
- Studies in interval support the measure
- There is good evidence for initial screening of patients with new onset diabetes for retinopathy, and ongoing eye exams in older adults with diabetes bi-annually if their diabetes is controlled And there is evidence of normal retinal examination in the past. If uncontrolled, or at high risk for retinal disease, and your examinations or more frequent examinations may be indicated. Evidence was updated in 2018.
- I believe we need a brief discussion for the following items:
- Should the denominator exclude patients with T1DM who were recently diagnosed as the recommendations are within 5 years of diagnosis? This is worth discussion.
- A reason for them to have an exam is of they they had very high glucose that was rapidly controlled and put them at risk for retinopathy.
- Will the numerator contain patients who received a retinal camera photo/exam?
- Overall, the quality of evidence supporting the guidelines and this measure is medium to strong, supporting the early identification of diabetic retinopathy and eye care to reduce visual impairments in diabetic patients. Some studies report a decline in diabetic retinopathy due to improvements in diabetic eye care and diabetic control. One study suggests that timely eye exam screenings and treatment in diabetics can prevent 75% of new blindness cases.
- The rational for this process of care measure is that adults with diabetes that do not receive regular retinal examinations are presumably at a higher risk for developing vision complications. There are at least two reasons for this. The first - blinding retinopathy may be asymptomatic and is treatable, so periodic screenning to detect problems before there is what may be irreversible vision loss is beneficial. Also, not really highlighted by the developer, since it has long been advocated by leading medical organizations that annual eye exams are appropriate for patients with diabetes meillitus, patients and their doctors who are not compliant with standard of care recommendations presumably are more likely to be non-compliant with diabetes related health care measures such as HbA1C control, and poor control is a key driver, with disease duration, of retinopathy. Based on the Evidence Algorithm on page 21, I agree the evidence is "moderate."

Performance Gap

- There is indeed a performance gap that warrants a national performance measure.
- NCQA provided HEDIS results for commercial, Medicare and Medicaid populations which showed that there was still room for improvement across populations. There were less than optimal outcomes across populations. The Developer provided performance data also from the 2015 PQRS reporting year with a mean of 78.1% and a standard deviation of 28.3%. The Developer provided performance data for the NCQA's Diabetes Recognition Program (DRP) from 2015, 2016, and 2017 with a mean: 61.4% (2014) to 62.8% (2016) and a standard deviation of 24.3% (2014) to 21.3% (2016). The developer did not provide disparities data. However they cited the 2014 study, using data from 2002-2009 found that that racial disparities and other influencing factors have an impact on rates of eye examinations among patients with diabetes and there needs to be more efforts to improve

screening and testing of diabetic retinopathy among minorities (Shi et al). They also provided the 2018 study, using data from 2013, noted that improvement in quality in diabetes care will help reduce diabetes complications and mortality (Canedo et al.).

- Among commercial providers percentage of diabetic patients receiving dilated retinal exams in the calendar year is in the low 50's. Medicare and Medicaid groups do better but the standard deviation for all groups is broad so there is variability in all groups.
- Blindness from Diabetes is largely preventable with proper care - there is great need for improvement.
- Data on the measure by population subgroups was not provided.
- The current measure performance is inconsistent across measured entities with some showing decreased performance.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- CMS QPP and HEDIS data support opportunity for improvement – high rating
- One study cited which demonstrates racial disparity on eye examinations in patients with diabetes
- Yes, there remains a gap that justifies the measure.
- Most recent performance data is 2016 for commercial Medicaid and Medicare And in the diabetes recognition program for 2017. There has not been much improvement in the mean rates of retinal examination's for any of these payer groups or programs. Mean completion rates are in the middle 50% range for commercial and Medicaid plans, 60% for the diabetes recognition program and nearly 70% for Medicare plans.
- Reporting does not allow stratification by demographic variables. Studies have shown that Hispanics, blacks and Asians had lower retinal examination rates than whites.
- There is a very poor national performance of this measure, with a mean of 61%. Means vary from commercial plans of 52% to Medicare performance of 68%
- Developer cited published material from the MEPS that stated racial disparities exist among other influencing factors
- HEDIS data from 2014-2016, data from NCQA's Diabetes Recognition Program from 2015-2017, and PQRS data from 2015 are provided. Comprehensive diabetes care remains a HEDIS measure for 2018. Data from the Medical Expenditure Panel Survey of 2014 found that racial disparities have an impact on rates of eye examinations among patients with diabetes.
- There is a clear performance gap. Hedis data shows 53% compliance (commercial), 54% compliance (Medicaid) and 68% compliance (Medicare). NCQA's Diabetes Recognition Program (DRP) performance was 61% and 2015 PQRS reporting year showed 78% compliance. Disparities -- Not really commented on in the document -- we know from many studies that underserved and poorer populations have less good control of their diabetes mellitus. Control is a key driver of retinopathy progression and severity. As is the case with much of medicine, there is a disparity gap here of course.

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

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

Evaluation of Reliability and Validity (and composite construction, if applicable): **Staff evaluation**

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] or [Scientific Methods Panel] is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

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

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

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

Committee pre-evaluation comments

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

Reliability
- No concerns with reliability - specifications. Da ta elements are clearly defined. No concerns with ability to consistently implement the measure.
- No issues with reliability of specifications
- Concur with the analysis of the staff evaluator.
- no concerns
- none
- Studies have shown that Hispanics, blacks and Asians had lower retinal examination rates than whites.
- No concerns about reliability. I do not think the measure needs voting for reliability.
- Reliability of the data in HEDIS and the Diabetic Recognition Program are discussed.
- No changes here that I note; no new concerns or comments.

Reliability Testing
- There are no reliability concerns.
- No concerns about the reliability of the measure.
- No concerns.
- Concur with the analysis of the staff evaluator.
- no concerns
- no

- No. This appears to have high reliability across the health plans.
- No concerns about reliability. I do not think the measure needs voting for reliability.
- No. Agree with "Moderate" rating

Validity Testing

- No concerns about validity of the measure/results. Performance scores are being compared and used for various programs including: health plan rankings and report cards, the State of Health Care Quality Annual Report, California Integrate Healthcare Association Pay for Performance program, Accountable Care Organization and Health Plan Accreditation programs, the Diabetes Recognition program and in Quality Compass.
- No concerns about validity testing. No threats to validity. This issue is finding a way to ensure that all (or most) individuals with diabetes have retinal exams on a yearly basis.
- Concur with the analysis of the staff evaluator.
- Staff analysis suggests that there are no exclusions in Potential Threats to validity however, hospice care is an exclusion but should not pose threat to validity;
- Rating Moderate. Therefore no need for discussion or vote
- Information regarding reasonable construct and face validity at both health plan and provider level are provided. These seem adequate.
- The interquartile range is statistically significant across all product lines and at the physician level. This suggests meaningful difference is in performance.
- If the questions regarding evidence (numerator and denominator) are clarified. I do not think the measure needs voting for validity.
- Validity of HEDIS measures is discussed.
- No. Agree with "Moderate" rating.

Other threats

- There are no significant threats to validity.
- No concerns with exclusions or risk adjustment in the specifications for the measure. The Developer did not consider social determinants or social risk factors in the measure development, though one could argue that these factors are visible in the outcome of measure reporting for commercial, Medicare and Medicaid HEDIS rates.
- No threats to validity from exclusions or risk adjustment
- Why is there a requirement for two outpatient visits versus one?
- N/A
- N/a
- If the questions regarding evidence (numerator and denominator) are clarified. I do not think the measure needs voting for validity.
- Exclusions - in hospice, gestational or steroid induced DM (Agree). Patients 65 and older with advanced illness condition and frailty (why have age criterion on this exclusion?)

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 measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records). While only some data elements are in defined fields in electronic sources, the elements are generated as byproduct of care processes. This measure is also a HEDIS measure and NCQA conducts audits to verify that HEDIS specifications are met.
- This is not an eMeasure.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- The data is routinely generated during routine care delivery.
- Uses claims and chart data. It is feasible but is more costly to capture data elements and more resource intensive. Should it become an e-measure and capture data out of electronic health records, costs may come down.
- These data elements are routinely generated and used during care delivery.
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- Data elements are routinely generated and used during care delivery
- Available in electronic form (not all data elements are in defined fields)
- Data collection already in operational use
- Rating of feasibility: moderate
- variability in data source for exam creates challenges in capturing data. Reports come via fax and paper charts, and often requires capture into structured data by office staff. This is additional staff work with little value add.
- Data from Eye examinations accomplished outside of the health system may not be within a providers chart as a discrete data element, making in accuracies for reported data from within an electronic health record. Claims data would most likely be available to the players who are being evaluated by this measure.
- The measure is very feasible to perform. It is easily measured through administrative/billing data. it is easy to measure if the patient was referred, but It is only measurable via EHR if distinct fields exist once a consult is received.
- The HEDIS Audit process is described.
- No new concerns. The measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records). While only some data elements are in defined fields in electronic sources, the elements are generated as byproduct of care processes. This measure is also a HEDIS measure and NCQA conducts audits to verify that HEDIS specifications are met. Agree with "modeerate" rating.

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

- HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
- STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
- INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States.
- ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.
- DIABETES RECOGNITION PROGRAM: This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care.
- QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial, Medicaid, and Medicare health plans.

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

- NCQA publishes HEDIS results annually in its Quality Compass tool. The measure receives feedback through the Policy Clarification Support System. The feedback received has generally been centered around clarification on which type of health care professional can review eye exams, types of photography that can count as an eye exam, and whether specific documentation counts as a negative or positive diagnosis for retinopathy.
- NCQA has provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

Additional Feedback: NA

***Questions for the Committee**:*

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

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

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

4b. Usability 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

- Overall, this measure has shown slight improvement for Medicare plans, a slight decline in performance for commercial plans, and a no change for Medicaid plans over the past three years. (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). These data are nationally representative.

- Since 2013, there has been an increase in the number of reporting physicians seeking recognition in NCQA's DRP program and an increase in performance, however from 2015-2017 there has been a slight decline in number of physicians and practices (see summary data in 1b.2.)

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

Unexpected findings (positive or negative) during implementation

- The developer did not report any unexpected findings.

Potential harms

- The developer did not report any unintended consequences.

Additional Feedback: NA

***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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- The data is used in public reporting and accountability programs.
- Used for health plan rankings and report cards, the State of Health Care Quality Annual Report, California Integrate Healthcare Association Pay for Performance program, Accountable Care Organization and Health Plan Accreditation programs, the Diabetes Recognition program and in Quality Compass. Feedback is built into the process to collect the data, including the HEDIS scores.
- The measure is widely used and reported - see Measure Maintenance Document. Feedback is provided regarding rates of perforance.
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity
- No concerns
    - Part of NCQA health plan report cards and publically reported;
    - Part of NCQA Health Care annual report
    - NCQA Accreditation for ACO and health plans
    - NCQA Diabetes recognition program for physicians

- This measures being publicly reported in many venues. Data is available for those being measured, and also to populations they serve.
- It is reported in 7 nationally recognized Quality Improvement systems, as well as required reporting for ACO accreditation.
- Most feedback is as questions for clarifications of items I asked in the Evidence section…concerning photos and supporting documents (consultation letters) and as I mentioned in the Feasibility section.
- HEDIS data is published in numerous publications and many types of providers reference HEDIS reports. Diabetes Recognition Program measurements are shared with the participating clinicians.
- Pass. The meausre is publicly reported is used in accountability programs. See the list starting at the bottom of page 6 and continuing to the top of page 7 of and of the measure worksheet. Results are disclosed.

Usability

- No concerns
- No identified harms. Usability is high as it reflects current standard of care by Associations of providers of care as well as the American Diabetes Association.
- There is room for improvement. When commercial insurance providers only obtain retinal exams on about 50% of diabetic patients much works is left to be done.
- There is no discussion of challenges to improvement – Measure Developer reports that performance is stable which does not mean improved.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- no concerns
    - Slight improvement for Medicare plans, decrease for commercial
    - From 2015 on there has been decrease in number of physicians in NCQA DRP

- Little if any improvement has been seen over the past 3 to 4 years in this measure. There is no foreseen unintended consequences or harms.
- Plans and practices that are aware of low performance rates usually become proactive to facilitate performance.
- There are no harms anticipated.
- By HEDIS data, this measure has shown slight improvement for Medicare plans, a slight decline in performance for commercial plans, and a no change for Medicaid plans over the past three years. The developer did not report any unintended consequences.
- Agree with the "Moderate" rating. Overall, this measure has shown slight improvement for Medicare plans, a slight decline in performance for commercial plans, and a no change for Medicaid plans over the past three years. Suggestion -- encourage more telemedicine use? It has been demonstrated in the literature that fundus photographs read by trained graders more reliably detects and categorizes retinopathy than office exams conducted by most providers. There is evidence telemediine approaches are cost effective. Evidence is accumulating that AI (artificial intelligence) approaches work, obviating the need to have or teach graders. Photograph codes do count in the numerator so such approaches are allowed by the measure; they could be encouraged more.

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0055
Measure Title: Comprehensive Diabetes Care: Eye exam (retinal) performed

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_0055" document

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

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

   ☒Yes (go to Question #2)

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

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

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

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

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   REFERENCE: testing attachment section 2b1.

   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

   ***TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*

   ☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:

    ☒High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

    ☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    REFERENCE: Testing attachment, section 2b2-2b6

    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

    ☒Yes (go to Question #12)

    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

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

☐Yes (please explain below then go to Question #13)

☐No (go to Question #13)

☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

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

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*

REFERENCE: *Testing attachment, section 2b1.*

TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 tesmting has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0055
Measure Title: Comprehensive Diabetes Care: Eye Exam (retinal) performed
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title
Date of Submission: 4/9/2018

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: Click here to name the health outcome

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

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: retinal exam in patients with diabetes

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

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Adults with diabetes (type 1 or 2) >>> Eye exam (retinal) is performed>>> Eye exam results are evaluated>>>Eye exam results are positive for diabetic retinopathy>>>Health provider determines treatment>>>Control of diabetic retinopathy and improvement in quality of life (desired outcome).

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.)

N/A

****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.
N/A

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

Table 1. American Diabetes Association (ADA) Guidelines

Source of Systematic Review: • **Title**	2018 Submission American Diabetes Association. (2018). Standards of Medical Care in Diabetes – 2018. Diabetes Care 2018; 41(Suppl. 1): S105-S118; doi: 10.2337/dc18-S010

• **Author** • **Date** • **Citation, including page number** • **URL**	Guideline available from: http://care.diabetesjournals.org/content/41/Supplement_1 2013 Submission American Diabetes Association. (2013). Standards of Medical Care in Diabetes – 2013. Diabetes Care 2013; 36:S1-e4; doi: 10.2337/dc13-S001 Guideline available from: http://care.diabetesjournals.org/content/36/Supplement_1/S11
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.	2018 Submission "Screening • Adults with type 1 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist within 5 years after the onset of diabetes (B) • Patients with type 2 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist at the time of the diabetes diagnosis (B) • If there is no evidence of retinopathy for one or more annual eye exam and glycemia is well controlled, then exams every 1–2 years may be considered. If any level of diabetic retinopathy is present, subsequent dilated retinal examinations should be repeated at least annually by an ophthalmologist or optometrist. If retinopathy is progressing or sight threatening, then examinations will be required more frequently. (B) • While retinal photography may serve as a screening tool for retinopathy, it is not a substitute for a comprehensive eye exam (E) • Women with preexisting type 1 or type 2 diabetes who are planning pregnancy or who are pregnant should be counseled on the risk of development and/or progression of diabetic retinopathy (B) • Eye examination should occur before pregnancy in the first trimester in patients with preexisting type 1 or type 2 diabetes, and then patients should be monitored every trimester and for 1 year postpartum as indicated by the degree of retinopathy (B) 2013 Submission "Screening • Adults and children aged ≥10 years with type 1 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist within 5 years after the onset of diabetes. (B) • Patients with type 2 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist shortly after the diagnosis of diabetes. (B) • Subsequent examinations for type 1 and type 2 diabetic patients should be repeated annually by an ophthalmologist or optometrist. Less frequent exams (every 2–3 years) may be considered following one or more normal eye exams. Examinations will be required more frequently if retinopathy is progressing.(B) • High-quality fundus photographs can detect most clinically significant diabetic retinopathy. Interpretation of the images should be performed by a trained eye care provider. While retinal photography may serve as a screening tool for retinopathy, it is not a substitute for a comprehensive eye

	exam, which should be performed at least initially and at intervals thereafter as recommended by an eye care professional. (E) • Women with pre-existing diabetes who are planning pregnancy or who have become pregnant should have a comprehensive eye examination and be counseled on the risk of development and/or progression of diabetic retinopathy. Eye examination should occur in the first trimester with close follow-up throughout pregnancy and for 1 year postpartum. (B)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Level of evidence and description: • B: Supportive evidence from well-conducted cohort studies, including: o Evidence from a well-conducted prospective cohort study or registry o Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study • E: Expert consensus or clinical experience 2013 Submission Same as above
Provide all other grades and definitions from the evidence grading system	2018 Submission Level of Evidence & Description: • A: Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted multicenter trial o Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted trial at one or more institutions o Evidence from a meta-analysis that incorporated quality ratings in the analysis • C Supportive evidence from poorly controlled or uncontrolled studies o Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results o Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) o Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation 2013 Submission Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above

	2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The ADA does not provide information on the systematic review conducted to support its 2018 or 2013 guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on two other systematic reviews that support the ADA's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 2. American Academy of Ophthalmology (AAO) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Academy of Ophthalmology Retina Panel. Preferred Practice Pattern Guidelines. Diabetic Retinopathy. American Academy of Ophthalmology. 2017. 1-63 URL: https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017 2013 Submission American Academy of Ophthalmology Retina Panel. Preferred Practice Pattern Guidelines. Diabetic Retinopathy. American Academy of Ophthalmology. 2008. 1-43 **URL**: http://www.eyenet.com.cn/upfiles/2015-06/20150626160757_6751.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	2018 Submission Pg. 10 "Examination • Visual acuity (III; Good; Strong) • Slit-lamp biomicroscopy (III; Good; Strong) • Intraocular pressure (IOP) (III; Good; Strong) • Gonioscopy before dilation, when indicated. Iris neovascularization is best recognized prior to dilation. When neovascularization of the iris is present or

suspected, or if the IOP is elevated, undilated gonioscopy can be used to detect neovascularization in the anterior chamber angle. (III; Good; Strong)

- Pupillary assessment for optic nerve dysfunction
- Thorough funduscopy including stereoscopic examination of the posterior pole (III; Good; Strong)
- Examination of the peripheral retina and vitreous (III; Good; Strong)

A dilated pupil is preferred to ensure optimal examination of the retina, because only 50% of eyes are correctly classified for the presence and severity of retinopathy through undilated pupils.

- Slit-lamp biomicroscopy is the recommended method to evaluate retinopathy in the posterior pole and midperipheral retina. (III; Good; Strong)
- Examination of the peripheral retina is best performed using indirect ophthalmoscopy or slit-lamp biomicroscopy. (III; Good; Strong)

Because treatment is effective in reducing the risk of visual loss, a detailed examination is indicated to assess for the following features that often lead to visual impairment:

- Macular edema (III; Good; Strong)
- Signs of severe NPDR (extensive retinal hemorrhages/microaneurysms, venous beading, and IRMA) (III; Good; Strong)
- Optic nerve head neovascularization and/or neovascularization elsewhere (III; Good; Strong)
- Vitreous or preretinal hemorrhage (III, Good; Strong)

Page 9 -Table 3 Recommended Eye Examination for Patients with Diabetes Mellitus and No Diabetic Retinopathy

Diabetes Type	Recommended Initial Evaluation	Recommended Follow up*
Type 1	5 years after diagnosis (II++;Good, Strong)	Yearly (III; Good; Strong)
Type 2	At time of diagnosis (II+; Good Strong)	Yearly (III; Good; Strong)
Pregnancy (type 1 or type 2)	Soon after conception and early in the first trimester (III; Good; Strong)	No retinopathy to mild or moderate NPDR: every 3-12 months (III, Good, Strong) Severe NPDR or worse: every 1-3 months (III, Good, Strong)

NPDR= nonproliferative diabetic retinopathy

*Abnormal findings may dictate more frequent follow up examinations"

2013 Submission

Pg. 23

"Examination

- Visual acuity (A: I)
- Slit-lamp biomicroscopy (A:III)
- Intraocular pressure (A:III)
- Gonioscopy when indicated (A:III)
- Dilated funduscopy including stereoscopic examination of posterior pole (A:I)
- Examination of the peripheral retina and vitreous (A:III)

<table>
<tr>
<td></td>
<td>

A dilated pupil is necessary to ensure optimal examination of the retina, because only 50% of eyes are correctly classified for the presence and severity of retinopathy through undilated pupils (A:I). Slit-lamp biomicroscopy with accessory lenses is the recommended method to evaluate retinopathy in the posterior pole and midperipheral retina (A:III). The examination of the peripheral retina is best performed with indirect ophthalmoscopy or with slit-lamp biomicroscopy, combined with a contact lens (A:III).

Table A2-1 Recommended Eye Examination Schedule for Patients with Diabetes Mellitus
<table>
<tr><th>Diabetes Type</th><th>Recommended Time of First Examination</th><th>Recommended Follow up*</th></tr>
<tr><td>Type 1</td><td>3-5 years after diagnosis (A:II)</td><td>Yearly (A:II)</td></tr>
<tr><td>Type 2</td><td>At time of diagnosis (A:II)</td><td>Yearly (A:II)</td></tr>
<tr><td>Prior to pregnancy (type 1 or type 2)</td><td>Prior to conception and early in the first trimester (A:I)</td><td>No retinopathy to mild or moderate NPDR: every 3-12 months (A:I)
Severe NPDR or worse: every 1-3 months (A:I)</td></tr>
</table>
NPDR= nonproliferative diabetic retinopathy
*Abnormal findings may dictate more frequent follow up examinations”
</td>
</tr>
<tr>
<td>Grade assigned to the evidence associated with the recommendation with the definition of the grade</td>
<td>
2018 Submission

To rate individual studies, a scale based on Scottish Intercollegiate Guideline Network (SIGN) is used. The definition and levels of evidence to rate individual studies are as follows:

II++ High-quality systematic reviews of case-control or cohort studies
High-quality case-control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
II+ Well-conducted case-control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
III Nonanalytic studies (e.g., case reports, case series)

Recommendations for care are formed based on the body of the evidence. The body of evidence quality ratings are defined by GRADE as follows:

Good quality — Further research is very unlikely to change our confidence in the estimate of effect

Key Recommendations for care are defined by GRADE as follows:

Strong recommendation — Used when the desirable effects of an intervention clearly outweigh the undesirable effects or clearly do not

2013 Submission
Care Process Ratings:
</td>
</tr>
</table>

	• Level A: Most important to the care process Strength of Evidence Ratings: • Level I: includes evidence from at least one properly conducted, well-designed, randomized controlled trial. It could include meta-analyses of randomized controlled trials • Level II: includes evidence obtained from the following: o Well-designed controlled trials without randomization o Well-designed cohort or case -control analytic studies, preferably from more than one center o Multiple-time series with or without the intervention • Level III: include evidence obtained from one of the following: o Descriptive studies o Case reports o Reports of expert committees/organizations
Provide all other grades and definitions from the evidence grading system	2018 Submission To rate individual studies, a scale based on Scottish Intercollegiate Guideline Network (SIGN) is used. The definition and levels of evidence to rate individual studies are as follows: I++ High-quality meta-analyses, systematic reviews of randomized controlled trials (RCTs), or RCTs with a very low risk of bias I+ Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias I- Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias II- Case-control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal Recommendations for care are formed based on the body of the evidence. The body of evidence quality ratings are defined by GRADE as follows: Moderate quality — Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate Insufficient quality — Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate Any estimate of effect is very uncertain Key Recommendations for care are defined by GRADE as follows: Discretionary recommendation — Used when the trade-offs are less certain—either because of low-quality evidence or because evidence suggests that desirable and undesirable effects are closely balanced 2013 Submission Care Process Ratings: • Level B: Moderately important to the care process • Level C: Relevant but not critical to the care process

Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AAO does not summarize the details of the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the AAO systematic review, we provide information on two other systematic reviews that support the AAO's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 3. American Geriatrics Society (AGS) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Geriatrics Society (AGS). 2013. Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. American Geriatrics Society Panel on the Care for Older Adults with Diabetes Mellitus. Journal of American Geriatric Society. 2013 November; 61 (11): 2020-2026. Doi:10.1111/jgs.12514 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064258/pdf/nihms583558.pdf 2013 Submission American Geriatrics Society (AGS). 2003. Guidelines for Improving the Care of the Older Person with Diabetes Mellitus. California Healthcare Foundation/American Geriatrics Society Panel on Improving Care for Elders with Diabetes. American Geriatrics Society. May 2013; 51, Suppl 5, JAGS URL:

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.	2018 Submission "1. Older adults with new-onset DM should have an initial screening dilated-eye examination with funduscopy performed by an eye care specialist." (Level I, Grade B) "2. Older adults with DM and who are at high risk for eye disease (symptoms of eye disease present; evidence of retinopathy, glaucoma, or cataracts on an initial dilated-eye examination or subsequent examinations during the prior 2 years; A1C ≥ 8.0%; type 1 DM; or blood pressure ≥ 140/80) on the prior examination should have a screening dilated-eye examination performed by an eye-care specialist with funduscopy training at least annually. Persons at lower risk or after one or more normal eye examinations may have a dilated-eye examination at least every 2 years." (Level II, Grade B) 2013 Submission Page S272 "1. The older adult who has new-onset DM should have an initial screening dilated-eye examination performed by an eye-care specialist with funduscopy training." (Level I, Grade B) "2. The older adult who has DM and who is at high risk for eye disease (symptoms of eye disease present; evidence of retinopathy, glaucoma, or cataracts on an initial dilated-eye examination or subsequent examinations during the prior 2 years; A1C ≥ 8.0%; type 1 DM; or blood pressure ≥ 140/80) on the prior examination should have a screening dilated-eye examination performed by an eye-care specialist with funduscopy training at least annually. Persons at lower risk may have a dilated-eye examination at least every 2 years." (Level III, Grade B)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Quality of Evidence • Level I: Evidence from at least one properly randomized controlled trial • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytical studies, from multiple time-series, or from dramatic results in uncontrolled experiments Strength of Evidence • B: Moderate evidence to support the use of a recommendation clinicians "should do this most of the time" 2013 Submission Quality of Evidence • Level I: Evidence from at least one properly randomized controlled trial • Level III: Evidence from respected authorities, based on clinical experience, descriptive studies, or reports of expert committee Strength of Evidence • B: Moderate evidence to support the use of a recommendation; clinicians should do this most of the time
Provide all other grades and definitions from the evidence grading system	2018 Submission Quality of Evidence • Level III: Evidence from respected authorities based on clinical experience, descriptive studies, or reports of expert committees Strength of Evidence • A: Good evidence to support the use of a recommendation; clinicians should do this all the time

	• C: Poor evidence to support or to reject the use of a recommendation; clinicians may or may not follow the recommendation • D: Moderate evidence against the use of a recommendation; clinicians should not do this • E: Good evidence against the use of a recommendation; clinicians should not do this 2013 Submission • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytic studies, or from multiple time-series studies, or from dramatic results in uncontrolled experiments • A: Good evidence to support the use of a recommendation; clinicians should do this all the time • C: Poor evidence to support or to reject the use of a recommendation; clinicians may or may not follow the recommendation • D: Moderate evidence against the use of a recommendation; clinicians should not do this • E: Good evidence against the use of a recommendation; clinicians should not do this
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AGS does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the AGS systematic review, we provide information on two other systematic reviews that support the AGS's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 4. Additional Systematic Reviews

Citations	AACE Diabetes Care Plan Guidelines. Endocrine Practice. 2011. Vol 17, Suppl 2: 1-53 **URL:** http://journals.aace.com/doi/abs/10.4158/EP.17.S2.1	Li R, Zhang P, Barker LE, Chowdhury FM, Zhang X. Cost-effectiveness of interventions to prevent and control diabetes mellitus: a systematic review. Diabetes Care. 2010. 33(8):1872-1894. **URL:** http://care.diabetesjournals.org/content/33/8/1872.full.pdf+html
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	The measure is based on guidelines and evidence that support the use of regular eye exams for individuals with diabetes. The evidence reviews for diabetic retinopathy describe a two-step approach to detect diabetic retinopathy early and delay visual impairments. Evidence includes recommendations for the timing of eye exams, appropriate eye tests, and appropriate providers for referrals.	
Grade assigned for the quoted evidence with definition of the grade	Numerical descriptor (evidence level) / Semantic descriptor (reference methodology) 1 Randomized controlled trials (RCT) 2 Meta-analysis of nonrandomized prospective or case-controlled trials (MNRCT) 2 Prospective cohort study (PCS) 3 Cross-sectional study (CSS) 3 Surveillance study (registries, surveys, epidemiologic study, retrospective chart review, mathematical modeling of database) (SS) 4 No evidence (theory, opinion, consensus, review, or preclinical study) (NE) 1=strong evidence; 2=intermediate evidence; 3=weak evidence; and 4=no evidence.	Randomized controlled trials in this review follow the American Diabetes Association guidelines. Per the ADA guidelines, grades assigned to the evidence varied from A - C. Level of Evidence & Description: A Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., the "all or none" rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis B Supportive evidence from well-conducted cohort studies, including: • Evidence from a well-conducted prospective cohort study or registry

		• Evidence from a well-conducted meta-analysis of cohort studies • Supportive evidence from a well-conducted case-control study C Supportive evidence from poorly controlled or uncontrolled studies, including: • Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) • Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation
Provide all other grades and associated definitions of the evidence in the grading system	2 Nonrandomized controlled trial (NRCT)	Level of Evidence & Description: E Expert consensus or clinical experience
What is the time period covered by the body of evidence?	1984-2006	1994-2003
Quantity and Quality of Body of Evidence	• Timing of Eye Exams: 1 Cross sectional study (CSS), 1 Surveillance study (SS), 1 Prospective cohort study (PCS) • Referral to appropriate providers: 1 Meta-analysis of nonrandomized case-controlled study (MNRCT) • Appropriate eye tests: 1 Surveillance study (SS) • Preventive diabetic retinopathy methods: 2 Randomized controlled trials (RCT), 2 Prospective cohort study (PCS) • Diabetic Retinopathy Treatments: 1 Randomized controlled trial (RCT), 2 Review/no evidence (NE)	The systematic review included six studies that examined the cost effectiveness of preventing eye complications in diabetics and treating retinopathy. Four of these studies focused on the timing of eye examinations (i.e. every 6 months, annually, every 2 years, every 3 years, etc.). These studies included a literature review, cross sectional, longitudinal, and epidemiological studies. One epidemiological study focused on the type of eye test and one randomized prospective clinical trial on the treatment of retinopathy.

What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting the guidelines and this measure is medium to strong. While there are seven studies that examine the timing of eye examinations, there were no RCTs for this area. More RCTs were available when examining referrals to providers.
Estimates of benefit and consistency across studies in body of evidence – what are the estimates of benefits?	The evidence supports the early identification of diabetic retinopathy and eye care to reduce visual impairments in diabetic patients. Early detection of retinopathy also improves the quality of life in diabetics and reduces financial burdens that stem from poor visual health. Some studies report a decline in diabetic retinopathy due to improvements in diabetic eye care and diabetic control. One study also suggests that timely eye exam screenings and treatment in diabetics can prevent 75% of new blindness cases.
What harms were studied and how do they affect the net benefit (benefits over harms)?	Overall, there are minimal harms associated with receiving dilated eye examinations. Minor discomforts may stem from having the eyes dilated. One additional harm may include the misclassification of the level of diabetic retinopathy due to possible false negative exam results. These harms can be mitigated with regular subsequent eye exams based on the guidelines. These potential harms do not outweigh the benefits of having regular eye examinations to provide early detection of diabetic retinopathy.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous studies have been conducted since the systematic reviews we cite in this table, none of which change the conclusion that routine eye exams for individuals with diabetes are appropriate. Below we list two additional studies that support this measure. Nathan DM, Bebu I, Hainsworth D, et al.; DCCT/EDIC Research Group. Frequency of evidence-based screening for retinopathy in type 1 diabetes. N Engl J Med 2017;376:1507–1516 Agardh E, Tababat-Khani P. Adopting 3-year screening intervals for sight-threatening retinal vascular lesions in type 2 diabetic subjects without retinopathy. Diabetes Care 2011;34:1318–1319

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.

N/A

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

N/A

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

N/A

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

N/A

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

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

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

nqf_evidence_0055_Eye_Exam_7.1.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.

Yes

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 promotes regular eye examinations in diabetic adults (ages 18-75). Diabetic retinopathy and vision loss are complications from diabetes. Adults with diabetes that do not receive regular retinal examinations are at a higher risk for developing these vision complications. Vision screenings are part of high quality care for patients with diabetes.

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

The following data are extracted from HEDIS data collection reflecting the most recent years of reporting for this measure. Performance data is summarized at the health plan level and summarized by number of plans reporting, mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year and product line (i.e. commercial, Medicare, and Medicaid).

Comprehensive Diabetes Care: Eye Exam (Retinal) Performed
*Higher score= better performance
N= Number of plans reporting

Commercial Rate

YEAR	N	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX
2016	411	50.5%	12.6%	19.8%	35.7%	41.6%	49.8%	55.0%	68.0%	87.8%
2015	418	50.4%	12.6%	14.3%	34.5%	41.6%	48.9%	58.4%	69.0%	86.5%
2014	391	52.6%	12.3%	25.1%	37.5%	44.5%	50.9%	60.6%	70.4%	86.3%

Medicaid Rate

YEAR	N	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX
2016	271	54.9%	11.7%	15.3%	39.6%	47.6%	55.2%	63.5%	68.2%	87.8%
2015	261	52.8%	12.6%	14.9%	36.6%	44.5%	53.7%	61.5%	68.1%	88.7%
2014	220	54.4%	11.6%	23.2%	38.8%	47.1%	54.8%	63.3%	67.8%	87.3%

Medicare Rate

YEAR	N	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX

2016 | 473 | 70.2%| 11.0%| 25.8%| 56.2%| 64.2%| 71.0%| 77.7%| 83.1%| 96.6%
2015| 460 | 68.7%| 11.2%| 19.0%| 54.3%| 62.0%| 69.0%| 76.9%| 82.4%| 93.3%
2014| 475| 68.5%| 11.5%| 14.1%| 55.0%| 61.3%| 69.2%| 76.6%| 82.0%| 97.1%

This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care (see full description of program in 4a1.1).

Diabetes Recognition Program -
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX|
2017|3771|62.8%|21.3%|0.00%|28.0%|48.0%|68.0%|82.0%|88.0%|100.00%
2016|4704|60.2%|22.8%|0.00%|28.0%|48.0%|64.0%|77.3%|85.3%|100.00%
2015|4989|61.4%|24.3%|0.00%|25.7%|44.0%|64.0%|80.0%|88.6%|100.00%

PQRS
The following PQRS performance data includes claims, registry, measures group, GPRO Web Interface/ACO, QCDR data for services performed from in 2015.

Mean: 78.1%
St dev: 28.3%

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.
N/A

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.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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
A cross-sectional study aimed to observe the impact of racial disparities and other influencing factors such as age, gender, education, and insurance on eye examination rates among adults 18-64 years of age with diabetes. The study used data from the Medical Expenditure Panel Survey (MEPS) Household Component including the Diabetes Care Survey between 2002-2009. Eye examination rates were compared each year between non-Hispanic whites and minorities which included, but not limited to, black, American Indian/Alaska native, Asian, native Hawaiian/Pacific Islander). Between 2002-2009 there were approximately a weighted 60 percent of non-Hispanic whites compared to a 40 percent of minorities. The study found that across all years of the study, minorities had consistently lower unadjusted eye examination rates compared to non-Hispanic whites. Between 2002-2009, the unadjusted rate for eye examinations for minorities dropped from 56 percent to 49 percent while rates for non-Hispanic whites increased from 56 percent to 59 percent. When assessing associations between other influencing factors such as age, the study found that adults 45 years and older were more likely to receive an eye examination compared to adults between 18-45 years of age. The study also found that for all years except 2007, having health insurance was associated with an increased rate of eye examinations. Overall, the study found that racial disparities and other influencing factors has an impact on rates of eye

examinations among patients with diabetes and there needs to be more efforts to improve screening and testing of diabetic retinopathy among minorities (Shi et al., 2014).

Another cross-sectional study also analyzed MEPS data from 2013 to assess racial and ethnic disparities in diabetes quality of care among adults with type II diabetes. The study controlled for health insurance status, poverty, and education and observed the difference in adherence to five diabetes quality of care recommendations (HbA1c twice yearly, yearly foot exam, dilated eye exam, blood cholesterol test, and flu vaccination. Among 65 percent of patients who received an eye exam, Hispanics, blacks, and Asians had lower rates compared to whites. Overall, the study noted that improvement in quality if diabetes care will help reduce diabetes complications and mortality (Canedo et al., 2018).

Shi Q, Zhao Y, Fonseca V, Krousel-Wood M, & Shi L. Racial Disparity of Eye Examinations Among the U.S. Working-Age Population With Diabetes: 2002-2009. 2014. Diabetes Care;37:1321-1328, doi: 10.2337/dc13-1038.

Canedo JR, Miller ST, Schlundt D, Fadden MK, Sanderson M. Racial/Ethnic Disparities in Diabetes Quality of Care: the Role of Healthcare Access and Socioeconomic Status. 2018. Journal of Racial Ethnic Health Disparities;5(1):7-14. doi: 10.1007/s40615-016-0335-8.

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

Measure Number (*if previously endorsed*)**:** 0055
Measure Title: Comprehensive Diabetes Care: Eye exam (retinal) performed
Date of Submission: 3/5/2018
Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

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

N/A

1.3. What are the dates of the data used in testing? 2010-2012

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

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

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

Health Plan Level
We calculated the measure score reliability and construct validity from HEDIS data that included 416 commercial health plans, 500 Medicare health plans, and 197 Medicaid health plans. The sample included all commercial, Medicare, and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size

Physician Level:
We also calculated measure score reliability from physician/practice level data from the NCQA Diabetes Recognition Program (DRP) that included 3676 physicians. Construct validity was calculated with data from a sample of 653 physicians/practices.

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*)

2012 data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

HEDIS Health Plan

Product type	**Number of plans**	**Median number of eligible patients per plan**

Commercial HMO	218	7,433
Commercial PPO	198	14,513
Medicaid HMO	194	3,114
Medicare HMO	349	4,134
Medicare PPO	151	4,110

NCQA's Diabetes Recognition Program currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes. Below is a description of the sample. It includes the number of physicians and practices reporting on this measure in the DRP program in 2012.

Physician Level

Analysis	Number of physicians	Median denominator size
Reliability	3,676	25
Construct Validity	653	25

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

Reliability:
Reliability of the health plan measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Reliability of the physician/practice level measure in the DRP was tested using a beta-binomial calculation. This analysis included the entire DRP sample (described above).

Validity:
Validity of the health plan measure was demonstrated through construct validity using the entire HEDIS data sample (described above) and through a systematic assessment of face validity with expert panels.

Validity was demonstrated through construct validity using data from a sample of 653 physicians/practices and through a systematic assessment of face validity with expert panels.

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

We did not analyze performance by social risk factors.

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*)

Reliability testing of performance measure score:

Reliability was estimated by using the beta-binomial model for the health plan measure and physician/practice level DRP measure. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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*)

Health Plan Level

Product Type	**Reliability per Beta Binomial Model**
Commercial	1.0
Medicare	0.99
Medicaid	0.96

Physician Level

Product Type	**Reliability per Beta Binomial Model**
Diabetes Recognition Program	0.80

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?*)

Health Plan Level

The values for the beta-binomial statistic across all product lines for the health plan level measure suggest the measure has high reliability.

Physician Level
The value for the beta-binomial statistic for the physician level measure suggests the measure has high reliability.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Method of Testing Construct Validity – Health Plan Level
We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality hypothesized to be related, which are listed below.

- HbA1c Testing
- Hemoglobin (HbA1c) Poor Control (>9.0%)
- Medical attention for nephropathy
- Hemoglobin (HbA1c) Control (<8%)

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of Testing Construct Validity – Physician Level
We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality in NCQA's Diabetes Recognition Program hypothesized to be related, which are listed below.

- Blood pressure control
- Foot exam
- Medical attention for nephropathy

We tested the correlations using the Pearson correlation test described above.

Method of Assessing Face Validity – Health Plan Level

We describe below NCQA's process for both measure development and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assure that the measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.
STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed periodically, based on changes in evidence and guidelines. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support (PCS) portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures. Over the past four years, NCQA has received and responded to an average of 39 inquiries per year on this measure.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2 and in other relevant NCQA programs.

Method of Assessing Face Validity - Physician Level

The physician level measure was tested for face validity with four panels of experts. The Diabetes Recognition Program (DRP) Advisory Committee included 7 experts in diabetes care including representation by clinicians, health plans, integrated health systems and research organizations; DMAP, CPM and the Clinical Programs Committee (CPC). NCQA's CPC's oversees the evolution of NCQA's recognition programs and related measures including the Diabetes Recognition Program, the Heart/Stroke Recognition Program, the Patient Centered Medical Home and Patient-Centered Specialty Practice Recognition Program, among others. The CPC includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 18 members. The CPC is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of clinical recognition programs. CPC members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel

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

Construct Validity – Health Plan Level

The results from construct validity testing of the health plan level measure are presented by product line in Tables 1a, 1b, and 1c below.

Table 1a. Correlations among Diabetes Measures in Commercial Health Plans - 2012

	Pearson Correlation Coefficient			
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Medical Attention for Diabetic Nephropathy	HbA1c Good Control (<8.0%)
CDC - Eye Exams	0.69	-0.63	0.72	0.65

Note: All correlations are significant at p<0.0001

Table 1b. Correlations among Diabetes Measures in Medicaid Health Plans - 2012

	Pearson Correlation Coefficient			
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Medical Attention for Diabetic Nephropathy	HbA1c Good Control (<8.0%)
CDC - Eye Exams	0.53	-0.53	0.45	0.51

Note: All correlations are significant at p<0.0001

Table 1c. Correlations among Diabetes Measures in Medicare Health Plans - 2012

	Pearson Correlation Coefficient			
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Medical Attention for Diabetic Nephropathy	HbA1c Good Control (<8.0%)
CDC - Eye Exams	0.60	-0.52	0.38	0.51

Note: All correlations are significant at p<0.0001

Construct Validity – Physician Level

Table 2a below provides the results from construct validity testing of the physician level measure.

Table 2a. Correlations among Diabetes Measures in the NCQA Diabetes Recognition Program - 2012

	Pearson Correlation Coefficient		
	Blood Pressure Control	Medical Attention to Nephropathy	Foot Exam
CDC – Eye Exams	0.41	0.26	0.42

Note: All correlations are significant at p<0.0001

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?*)

Construct Validity – Health Plan Level

Across all product lines, the correlations are moderate to strong and statistically significant. These results confirmed the hypothesis that the diabetes measures are correlated with each other. Coefficients with absolute value of less than .3 are generally considered indicative of weak associations. Absolute values of .3 to .59 are considered moderate associations, absolute values of .6 to .69 indicate a strong positive relationship, and absolute values of .7 or higher indicate a very strong positive relationship. These correlation results suggest that at the plan level the measure has sufficient validity.

Note: Correlation values with the HbA1c Poor Control measure are all negative because it is a "lower is better quality" measure, while the other measures are all "higher is better".

Construct Validity - Physician Level

At the physician level, the *CDC-Eye Exam* measure has a moderate correlation with the *Blood Pressure Control* and *Foot Exam* measures in the Diabetes Recognition Program. The correlation between the *Eye Exam* measure and the *Medical Attention to Nephropathy* measure is lower and indicates a slightly weaker association. Overall these correlation results suggest that the physician level measure has sufficient validity.

Face Validity – Health Plan Level

NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that the *CDC – Eye Exams* measure is measuring what it intends to measure. The results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

Face Validity – Physician Level

The results indicate that the multiple experts, stakeholders and NCQA's Clinical Programs Committee concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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*)

Testing was not performed for the excluded sample.

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*)

Testing was not performed for the excluded sample.

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*)

Testing was not performed for the excluded sample.

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

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

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

☒ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

N/A

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

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**
☐ **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*)
N/A

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method

calculates a testing statistic based on the sample, size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans performance is significantly different from each other.

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*)

Health Plan Level - 2012

Product Type	N	Mean (%)	St Dev (%)	P10th (%)	P25th (%)	P50th (%)	P75th (%)	P90th (%)	IQR (%)	P value
Commercial HMO	218	56.82	13.83	38.44	47.69	57.58	66.83	75.13	19.14	<0.05
Commercial PPO	198	48.80	10.03	33.20	43.08	49.85	54.74	60.25	11.66	<0.05
Medicaid HMO	194	53.22	11.96	37.14	44.37	54.43	62.46	67.64	18.09	<0.05
Medicare HMO	349	66.79	11.40	53.04	59.85	67.35	74.35	80.87	14.60	<0.05
Medicare PPO	151	64.63	10.34	53.04	56.94	64.72	70.56	77.86	13.62	<0.05

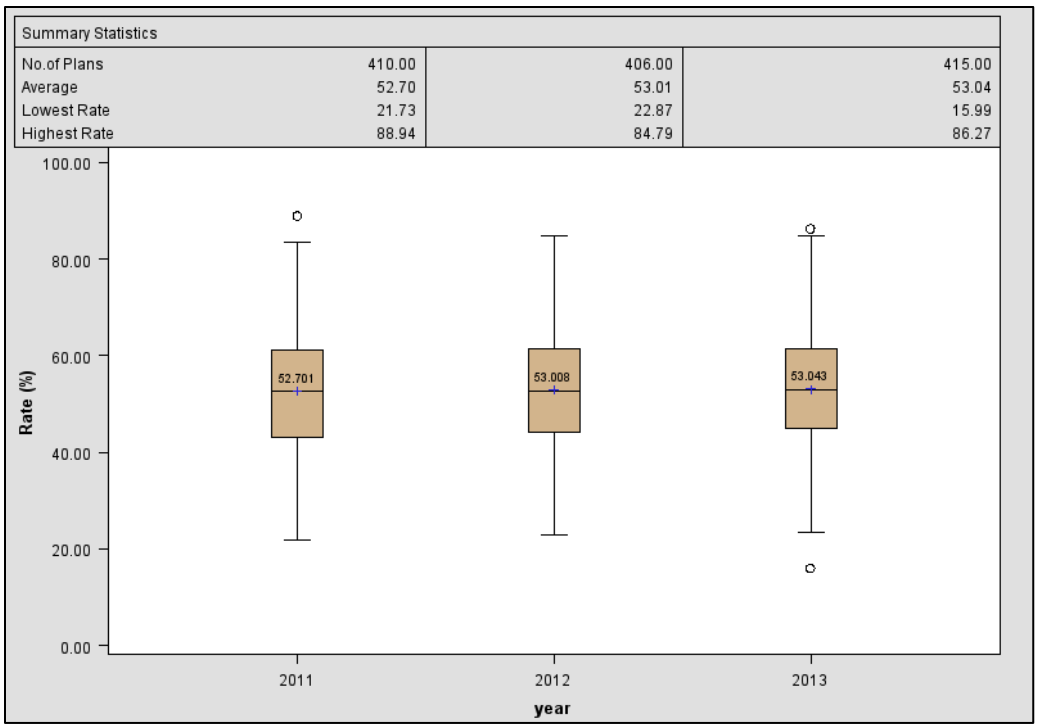
N = total number of plans reporting data
IQR: Interquartile range
p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

Physician Level - 2012

N (# of clinicians)	Mean (%)	St Dev (%)	10th (%)	25th (%)	50th (%)	75th (%)	90th (%)	IQR (%)	p value
3676	52.0	25.6	16.0	32.0	53.0	72.0	87.0	40.0	<0.05

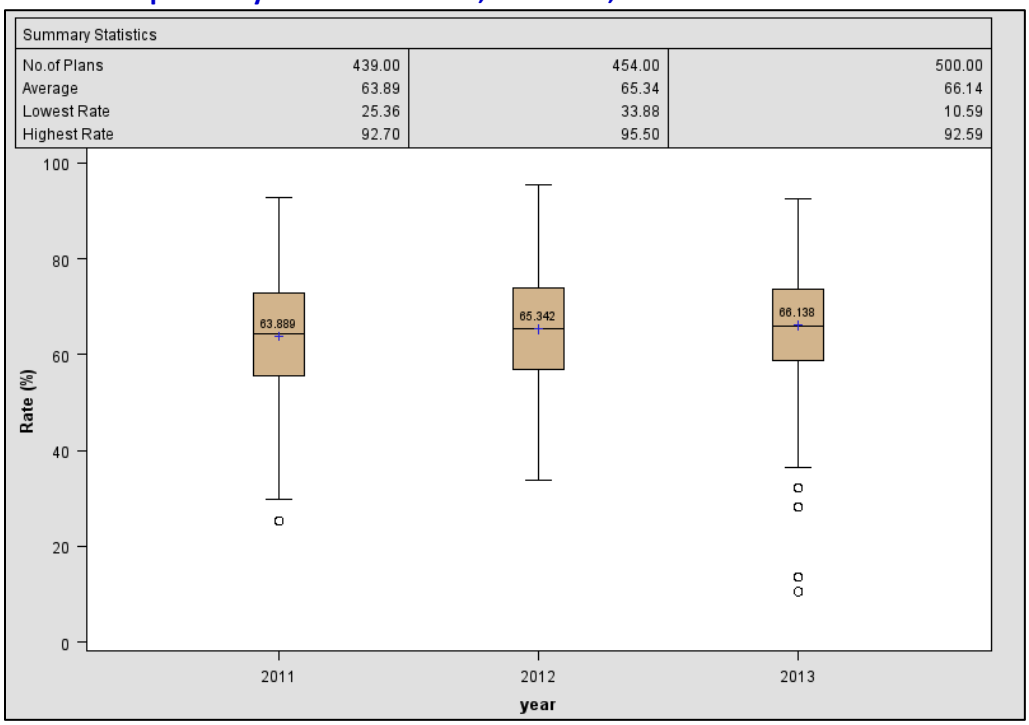
IQR: Interquartile range
p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile
Health Plan

Chart 1. Boxplot of Eye Exams Measure, Commercial, HEDIS 2011-2013*



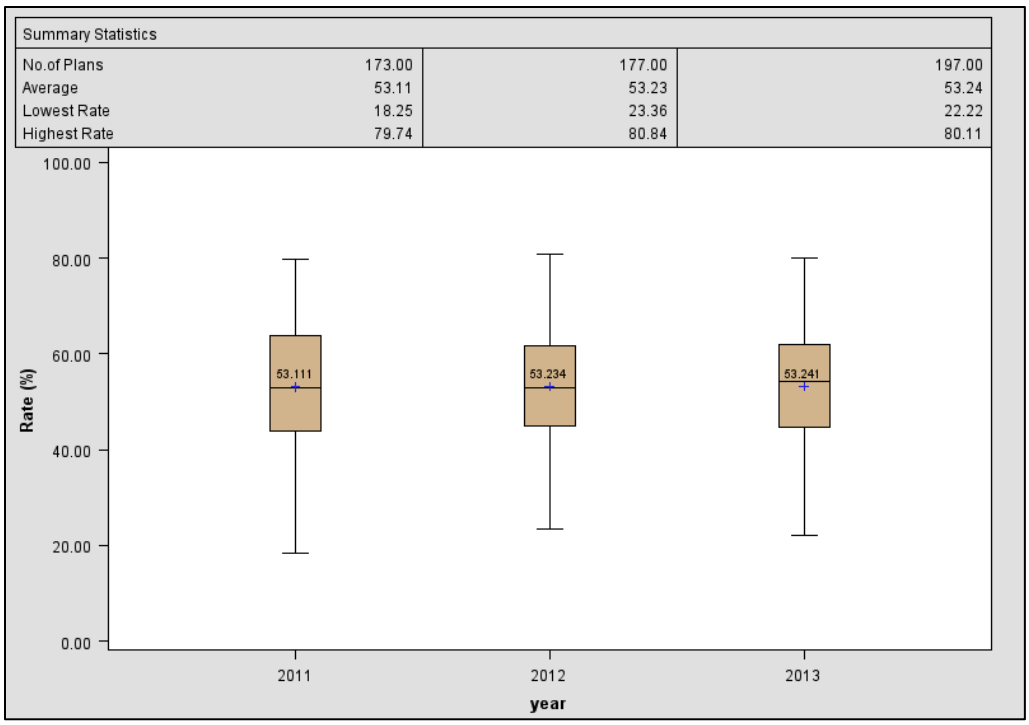
* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Chart 2. Boxplot of Eye Exams Measure, Medicare, HEDIS 2011-2013*



* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

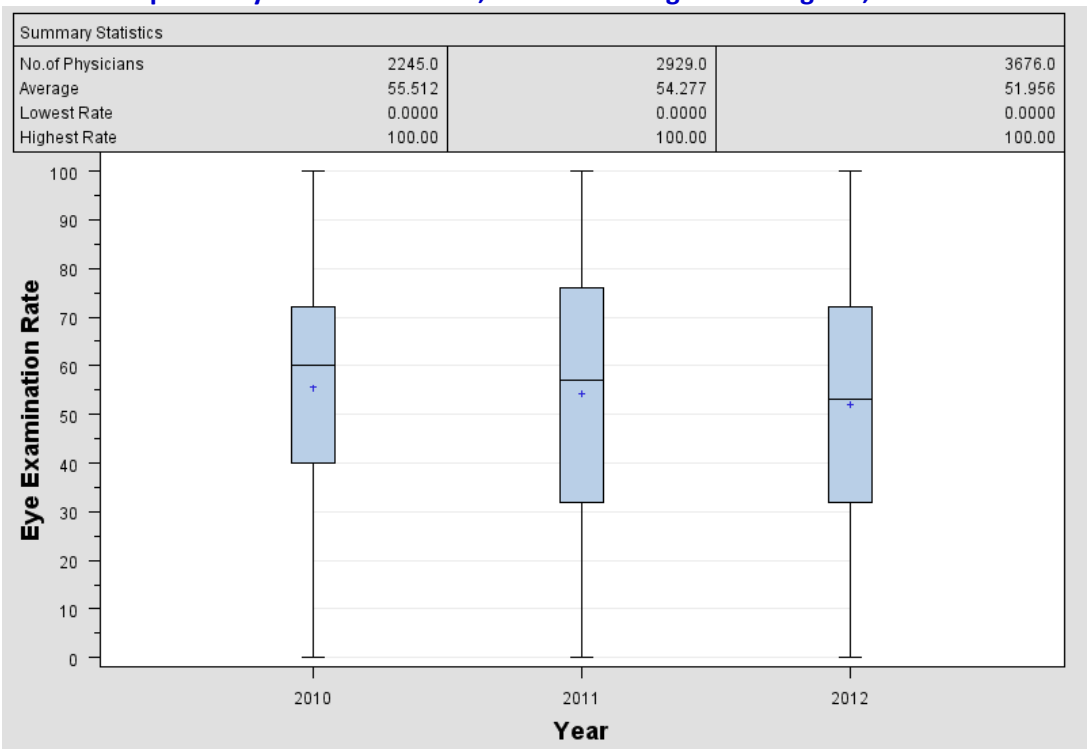
Chart 3. Boxplot of Eye Exams Measure, Medicaid HEDIS 2011-2013*



* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Physician Level

Chart 4. Boxplot of Eye Exams Measure, Diabetes Recognition Program, 2010-2012



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?*)

Health Plan Level

Across all product lines, the difference between the 25th (better performance) and 75th percentile is statistically significant. Overall, these results suggest there are meaningful differences in performance.

Physician Level
The difference between the 25th and 75th percentile is statistically significant, suggesting there are meaningful differences in performance.

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

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*)

This measure is collected with a complete sample.

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*)

This measure is collected with a complete sample.

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*)

This measure is collected with a complete sample.

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):*
Endocrine, Endocrine : Diabetes

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

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

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.)*
NA

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0055_CDC_Eye_Exam_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

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

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
-An additional route for numerator compliance was added to the measure which includes: Bilateral eye enucleation anytime during the patient's history through December 31 of the measurement year. This was added because these patients do not have retina's to examine
-Added another optional exclusion which is to exclude patients 65 and older with an advanced illness condition and frailty. This was added because quality measures that were intended for the general population may not be clinically appropriate or priority for individuals with advanced illness.

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

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who received an eye screening for diabetic retinal disease. This includes people with diabetes who had the following:
-a retinal or dilated eye exam by an eye care professional (optometrists or ophthalmologist) in the measurement year
–a negative retinal exam or dilated eye exam (negative for retinopathy) by an eye care professional in the year prior to the measurement year.
-Bilateral eye enucleation anytime during the patient's history through December 31 of the measurement year

For exams performed in the year prior to the measurement year, a result must be available.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Time period for data: a measurement year (12 months)

ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying numerator events for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD: At a minimum, documentation in the medical record must include one of the following:

- A note or letter prepared by an ophthalmologist, optometrist, PCP or other health care professional indicating that an ophthalmoscopic exam was completed by an eye care professional (optometrist or ophthalmologist), the date when the procedure was performed and the results.
- A chart or photograph indicating the date when the fundus photography was performed and evidence that an eye care professional (optometrist or ophthalmologist) reviewed the results. Alternatively, results may be read by a qualified reading center that operates under the direction of a medical director who is a retinal specialist.
-Evidence that the member had bilateral eye enucleation or acquired absence of both eyes. Look as far back as possible in the member's history through December 31 of the measurement year.
-Documentation of a negative retinal or dilated exam by an eye care professional (optometrist or ophthalmologist) in the year prior to the measurement year, where results indicate retinopathy was not present (e.g., documentation of normal findings). Documentation does not have to state specifically "no diabetic retinopathy" to be considered negative for retinopathy; however, it must be clear that the patient had a dilated or retinal eye exam by an eye care professional (optometrist or ophthalmologist) and that retinopathy was not present. Notation limited to a statement that indicates "diabetes without complications" does not meet criteria.

The patient is numerator compliant if the eye exam was performed in the measurement year or a negative eye exam was documented in the year prior to the measurement year. The patient is not numerator compliant if the eye exam or negative result are missing. Ranges and thresholds do not meet criteria for this measure.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients with diabetes can be identified two ways:
-CLAIM/ENCOUNTER DATA: Patients who had two face-to-face encounters, in an outpatient setting, observations visits, ED setting on different dates of service, or nonacute inpatient setting with a diagnosis of diabetes, or one face-to-face encounter in an acute inpatient, with a diagnosis of diabetes, during the measurement year or the year prior to the measurement year. Organizations may count services that occur over both years.

*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

-PHARMACY DATA: Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year.

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES (TABLE CDC-A):
Alpha-glucosidase inhibitors:
Acarbose, Miglitol

Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metaformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metaformin-saxagliptin, Metformin-sitagliptin , Sitagliptin-simvastatin

Insulin:
Insulin aspart, Insulin aspart-insulin aspart protamine, insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitagliptin

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Exclude patients who did not have a diagnosis of diabetes, in any setting, AND who had a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year
-Exclude patients 65 and older with an advanced illness condition and frailty

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.)*
ADMINISTRATIVE CLAIMS:
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD:
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and had a diagnosis of polycystic ovaries any time in the patient's history through December 31 of the measurement year.
OR
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year.

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

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 = Higher 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.)*
STEP 1. Determine the eligible population. To do so, identify patients who meet all the specified criteria.
-AGES: 18-75 years as of December 31 of the measurement year.
-EVENT/DIAGNOSIS: Identify patients with diabetes in two ways: by claim/encounter data and by pharmacy data.
Claim/Encounter Data:
-Patients who had at least two outpatient visits, observation visits, ED visits or nonacute inpatient encounters on different dates of service, with a diagnosis of diabetes. Visit type need not be the same for the two visits.
-Patients with at least one acute inpatient encounter with a diagnosis of diabetes.
*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

Pharmacy Data:
Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year.
*SEE PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES IN QUESTION S.7

STEP 2. Determine the number of patients in the eligible population who had a recent eye exam (retinal) performed during the measurement year through the search of administrative data systems.
STEP 3. Identify patients with a most recent eye exam (retinal) performed and the result.
STEP 4. Identify the most recent eye exam (retinal) during the measurement year or a negative result prior to the measurement year (numerator compliant). Identify missing eye exam or missing eye exam result (not numerator compliant).
STEP 5. Exclude from the eligible population patients from step 2 for whom administrative system data identified an exclusion to the service/procedure being measured.
*SEE DENOMINATOR EXCLUSION CRITERIA IN QUESTION S.8
STEP 6. Calculate the rate (number of patients with an eye exam (retinal) performed during the measurement year or negative result prior to the measurement year).

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Paper Medical 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.
This measure uses a combination of administrative claims data and medical records. Eye screening for diabetic retinal disease can be identified by the following administrative data:
-Retinal or dilated eye exam by an eye care professional (optometrist or ophthalmologist) in the measurement year.
-A negative retinal or dilated eye exam (negative for retinopathy) by an eye care professional in the year prior to the measurement year.
-Bilateral eye enucleation anytime during the patient's history through December 31 of the measurement year

Codes in the following value sets will meet these criteria:
-Any code in the Diabetic Retinal Screening Value Set billed by an eye care professional (optometrist or ophthalmologist) during the measurement year.
-Any code in the Diabetic Retinal Screening Value Set billed by an eye care professional during the year prior to the measurement year, with a negative result (negative for retinopathy).
-Any code in the Diabetic Retinal Screening Value Set billed by an eye care professional (optometrist or ophthalmologist) during the year prior to the measurement year, with a diagnosis of diabetes without complications
-Any code in the Diabetic Retinal Screening with Eye Care Professional Value Set billed by any provider type during the measurement year.
-Any code in the Diabetic Retinal Screening with Eye Care Professional Value Set billed by any provider type during the year prior to the measurement year, with a negative result (negative for retinopathy).
-Any code in the Diabetic Retinal Screening Negative Value Set billed by any provider type during the measurement year.
-Unilateral eye enucleation (Unilateral Eye Enucleation Value Set) with a bilateral modifier (Bilateral Modifer Value Set)
-Two unilateral eye enucleations (Unilateral Eye Enucleation Left Value Set) with service dates 14 days or more part.
-Left unilateral eye enucleation (Unilateral Eye Enucleation Left Value Set) and right unilateral eye enucleation (Unilateral Eye Enucleation Right Value Set) on the same or different dates of service

The minimum medical record documentation includes one of the following:
- A note or letter prepared by an ophthalmologist, optometrist, PCP or other health care professional indicating that an ophthalmoscopic exam was completed by an eye care professional (optometrist or ophthalmologist), the date when the procedure was performed and the results.
- A chart or photograph indicating the date when the fundus photography was performed and evidence that an eye care professional (optometrist or ophthalmologist) reviewed the results. Alternatively, results may be read by a qualified reading center that operates under the direction of a medical director who is a retinal specialist.
-Evidence that the member had bilateral eye enucleation or acquired absence of both eyes. Look as far back as possible in the member's history through December 31 of the measurement year.
-Documentation of a negative retinal or dilated exam by an eye care professional (optometrist or ophthalmologist) in the year prior to the measurement year, where results indicate retinopathy was not present (e.g., documentation of normal findings). Documentation does not have to state specifically "no diabetic retinopathy" to be considered negative for retinopathy; however, it must be clear that the patient had a dilated or retinal eye exam by an eye care professional (optometrist or ophthalmologist) and that retinopathy was not present. Notation limited to a statement that indicates "diabetes without complications" does not meet criteria.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Health Plan

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

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

2. Validity – See attached Measure Testing Submission Form
nqf_testing_0055_Eye_Exam_7.1_updated_4.18.18.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
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**.
Some data elements are in defined fields in electronic sources

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).
To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

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

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.

Broad public use and dissemination of these measures are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

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 Health Plan Rating http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx QPP https://qpp.cms.gov/ Health Plan Rating http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx QPP https://qpp.cms.gov/ Payment Program IHA California Pay for Performance IHA: http://www.iha.org/manuals_operations_2014.html Medicare Advantage Plan Rating https://www.medicare.gov/find-a-plan/questions/home.aspx Regulatory and Accreditation Programs NCQA Accreditation http://www.ncqa.org/Programs/ Accountable Care Organizations (ACO) http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx NCQA Accreditation http://www.ncqa.org/Programs/ Accountable Care Organizations (ACO)

	http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Professional Certification or Recognition Program Diabetes Recognition Program http://www.ncqa.org/Programs/Recognition/DiabetesRecognitionProgramDRP.aspx Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

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

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

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population

HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

MEDICARE ADVANTAGE PLAN RATING: This measure is included in the composite Medicare Advantage Star Rating. CMS calculates a Star Rating (1-5) for all Medicare Advantage health plans based on 53 performance measures. Medicare beneficiaries can view the star rating and individual measure scores on the CMS Plan Compare website. The Star Rating is also used to calculate bonus payments to health plans with excellent performance. The Medicare Advantage Plan Rating program covers 11.5 million Medicare beneficiaries in 455 health plans across all 50 states.

CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial, Medicaid, and Medicare health plans. As of Fall 2017, a total of 184 Medicare Advantage health plans were accredited using this measure among others covering 9.2 million Medicare beneficiaries; 451 commercial health plans covering 113 million lives; and 125 Medicaid health plans covering 35 million lives. Health plans are scored based on performance compared to benchmarks.

DIABETES RECOGNITION PROGRAM: This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care. The program

currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. The DRP Program has 6 measures which cover other areas such as: HbA1c control, blood Pressure control, eye examinations, nephropathy assessment, smoking and tobacco use and cessation advice or treatment, and foot examinations. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes.

INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of eight health plans representing 10 million insured persons. IHA is responsible for collecting data, deploying a common measure set, and reporting results for approximately 35,000 physicians in nearly 200 physician groups. This program represents the longest running U.S. example of data aggregation and standardized results reporting across diverse regions and multiple health plans. California consumers benefit from the availability of standardized performance results from a common measure set, which are available to the public through the State of California, Office of the Patient Advocate

QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible clinicals (ECs).

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?*)
N/A

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.*)
N/A

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

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

Questions received through the Policy Clarification Support system have generally centered around clarification on which type of health care professional can conduct and review eye exams, types of photography that can count as an eye exam, and whether specific documentation counts as a negative or positive diagnosis for retinopathy

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the PQRS and the Qualified Health Plan Quality Rating System.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
We have provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

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.
Overall, this measure has shown slight improvement for Medicare plans, a slight decline in performance for commercial plans, and a no change for Medicaid plans over the past three years. (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). These data are nationally representative.

Since 2013, there has been an increase in the number of reporting physicians seeking recognition in NCQA's DRP program and an increase in performance, however from 2015-2017 there has been a slight decline in number of physicians and practices (see summary data in 1b.2.)

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.**
There were no identified unexpected findings during testing or since implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits during testing or since implementation of this measure.

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.
No

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?
No

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
DIABETES EXPERT PANEL:
Bill Herman (Chair), MD, Univ. of Michigan Health System
David Aron, MD, Department of Veteran's Affairs
James Fain, PhD, RN, University of Massachussetts
Jerry Cavallerano, OD, Beetham Eye Institute
John Thompson, MD, Retina Specialists
Judith Fradkin, MD, NIDDK/NIH
Lynne Levitsky, MD, Massachusetts General Hospital

Mark Cziraky, PharmD, Healthcore
Richard Hellman, MD, Private Practice, Diabetes & Endocrinology
Seth Rubenstein, DPM, Reston Hospital Center, INOVA Fair Oaks Hospital
Stephen Fadem, MD, Baylor College of Medicine
Ted Ganiats, MD, Univ. of California, San Diego
Nancy Van Vessem, MD, Capital Health Plan

HEDIS EXPERT CODING PANEL
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Aqurate Health Data Management, Inc.
Nelly Leon-Chisen, RHIA, American Hospital Association
Alec McLure, MPH, RHIA, CCS-P, Verscend Technologies
Michele Mouradian, RN, BSN, Change HealthCare
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, WellCare Health Plans, Inc.

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

CLINICAL PROGRAMS COMMITTEE
Randall Curnow, MD, MBA, FACP, FACHE, FACPE (Chair), TriHealth
Suzanne Berman, MD, FAAP, Plateau Pediatrics
Brooks Daveman, MPP, Tennessee Division of Health Care Finance and Administration
Marcus Friedrich, MD, MBA, FACP, New York State Department Health Empire State Plaza, Coming Towne
Jennifer Gutzmore, MD, Cigna
Melissa Hogan, MPH, Aon
Adriana Matiz, MD, FAAP, Ambulatory Care Network
Lisa Morrise, Marts, LAM Professional Services, LLC
Deborah Murph, MBA, BSN, RN, Cherokee Health Systems
Amy Nguyen Howell, MD, MBA, CAPG
Marc Rivo, MD, Population Health Innovations
Julie Schilz, BSN, MBA, Anthem
Pamela Slaven-Lee, DNP, FNP-C, CHSE, The George Washington University School of Nnursing
Lina Walker, PhD, AARP

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 1999

Ad.3 **Month and Year of most recent revision:** 12, 2013
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested
for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that
noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the
measure.
These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds
a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to
alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial
purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA. © 2018 by the National Committee for Quality Assurance.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0056
Measure Title: Diabetes: Foot Exam
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a foot exam (visual inspection and sensory exam with mono filament and a pulse exam) during the measurement year.
Developer Rationale: This measure promotes regular foot examinations in adults with diabetes (ages 18-75). Because of macrovascular compromise leading to arterial insufficiency and microvascular effects on nerve function, surveillance of skin integrity is very important for patients with diabetes. Poor foot care can lead to infections and ultimately amputations of the toe, foot, lower limb, or upper limb. As a result of amputations, patients often experience drastic declines in quality of life. In order to maintain optimal quality of life for persons with diabetes, it is vital to maintain the highest quality of foot care in diabetic populations.

Numerator Statement: Patients who received a foot exam (visual inspection and sensory exam with monofilament and pulse exam) during the measurement period.
Denominator Statement: Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year.
Denominator Exclusions:
-Patients with a diagnosis of secondary diabetes due to another condition (e.g. a diagnosis of gestational or steroid-induced diabetes)
-Patients who have had either a bilateral amputation above or below the knee, or both a left and right amputation above or below the knee before or during the measurement period.
-Exclude patients who were in hospice care during the measurement year

Measure Type: Process
Data Source: Claims, Electronic Health Data, Other, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- The developer briefly described the link between foot exam and the patient's health outcomes in reducing/improvement in diabetes complications and quality of life.
- The developer provided an updated clinical guideline from the American Diabetes Association (ADA) (2018) including recommendations for the following:
    - Clinicians should perform a comprehensive foot evaluation at least annually to identify risk factors for ulcers and amputations. **B grading**
    - The examination should include inspection of the skin, assessment of foot deformities, neurological assessment (10-g monofilament testing with at least one other assessment: pinprick, temperature, vibration), and vascular assessment including pulses in the legs and feet. **B grading**
    - The Level of Evidence grading was B and C for the recommendations provided by developer. B level recommendations used supportive evidence from well-conducted cohort studies. C level recommendations used supportive evidence from a well conducted case-control study
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guidelines, as this information was not available in the guidelines.
- The developer provided a clinical guideline from the American Geriatrics Society (AGS) (2013) including recommendation for the following:
    - Older adults with DM should have a careful foot examination at least annually to check skin integrity and to determine whether there is loss of sensation or decreased perfusion and more frequently if there is evidence of any of these findings **(IIIA)**
    - Quality of Evidence-Level III (definition): Evidence from respected authorities based on clinical experience, descriptive studies, or reports of expert committees
    - Strength of Evidence-A (definition): Good evidence to support the use of a recommendation; clinicians should do this all the time
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guideline, as this information was not available.
- The developer also provided an additional systematic review from the The Journal of the American Medical Association (2005) which focused on efficacy of methods advocated for preventing diabetic foot ulcers. No grading provided.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided updated guidelines from 2018 from the American Diabetes Association which continues to support their measure focus.
- The developer also provided an additional systematic review from the The Journal of the American Medical Association (2005) which focused on efficacy of methods advocated for preventing diabetic foot ulcers and provides details on Quantity, Quality, Consistency of the measure focus.

Exception to evidence

NA

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm
Process measure with systematic review (Box 3) ->Summary of the QQC provided (Box 4) ->Systematic review concludes moderate quality evidence.

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

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

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

Performance Data:

- Developer provided performance data for the NCQA's Diabetes Recognition Program (DRP) from 2015, 2016, and 2017. The mean ranged from 71.7% to 75.2%
- Developer provided performance data also from the 2015 PQRS reporting year with a mean of 56.3%.

Disparities:

- Developer did not provide disparities from the measure. However cited CDC data from 2010 that examined diabetic adults that received a foot exam in a given year. The CDC data was categorized based on race/ethnicity, age, sex, and education level.
  - In 2010, Hispanics had the lowest percentage of foot exams (59%) in comparison to Whites (71%) and Blacks (77%).
  - In the same year, smaller disparities were seen according to age. Nearly 75% of all adults with diabetes between ages 65-74 received a foot exam, about 73% of adults between ages 45-64 and 71.5% of adults over age 75.
  - There were not significant disparities by gender: In 2010, 72.3% of males and 70.7% of females received foot exams.
  - Adults with an education greater than high school received foot exams at 70% while adults with only a high school education received foot exams at 67.8%; this gap widens for adults that achieved less than a high school education with only 59.1% receiving foot exams.

Questions for the Committee:

- *Specific questions on information provided for gap in care.*
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The updated evidence supporting the measure is directionally the same and is stronger. With that in mind, there is no need to repeat the discussion and to vote on the evidence.
- Based on the evidence provided, there is no need to review and vote on the evidence.
- No new or additional studies other than those identified in the submission which included the American Diabetes Association, Geriatric Society, and the Journal of the American Medical Association.

- The Journal of the American Medical Association (2005) focused on efficacy of methods advocated for preventing diabetic foot ulcers and provides details on Quantity, Quality, Consistency of the measure focus.
- For structure, process, and intermediate outcome measures:
    - What is the relationship of this measure to patient outcomes? Logic is valid from patient care standpoint
    - How strong is the evidence for this relationship? Grade B
    - Is the evidence directly applicable to the process of care being measured? Yes
    - Quality of evidence from Algorithm is Moderate

- Updated evidence base – primarily B-Grade
- No concerns, no need for repeat discussion and vote
- Relationship of measure to pt outcomes: see rationale
- Strength of evidence: moderate
- Evidence applicable to the process of care being measured: yes

- Since the updated information is stronger and directionally the same, I agree that we do not nee to discuss an vote on the Evidence Section

- This measure is crucial to patient outcomes

- This is a process measure of the percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a foot exam (visual inspection and sensory exam with mono filament and a pulse exam) during the measurement year. These patients had a diagnosis of type 1 or type 2 diabetes during the measurement year. Those diagnosed with gestational diabetes, steroid-induced diabetes, persons utilizing hospice care, and those with bilateral leg amputations were excluded. The American Diabetes Association guideline from 2018 suggests performing a comprehensive foot exam at least annually that includes evaluation of neurological and vascular function along with inspections at every clinical visit. Those with abnormal pulses or claudication symptoms should have an ankle brachial index. Those who smoke, have loss of protective sensation or who have peripheral arterial disease should be referred to foot care specialists. The American Geriatrics Society guideline for improving diabetes control of 11/13 suggests that older adults with diabetes have a careful foot exam at least annually. A JAMA article from 2005 gives efficacy evidence for the examination techniques of the feet to help prevent diabetic foot ulcers. Two more recent articles support the JAMA article's conclusions.
- No new evidence that is not cited.
- This process measure is updated on the 2018 ADA standards of medical care in diabetes. Rating of evidence is moderate. Measure remains essentially on changed from previous review.

Performance Gap

- Based on the performance data presented, a performance gap continues to exist. Additionally, the evidence provided suggests that there are racial/ethnicity disparities that exist.
- There continues to be a significant performance gap on this measure.
- The Developer provided performance data for the NCQA's Diabetes Recognition Program (DRP) from 2015, 2016, and 2017. The mean ranged from 71.7% to 75.2%. The Developer provided performance data also from the 2015 PQRS reporting year with a mean of 56.3%. The results show a continued opportunity for improvement. The Developer did not address disparities in its submission. They did provide the following information: The CDC data was categorized based on race/ethnicity, age, sex, and education level.
    - In 2010, Hispanics had the lowest percentage of foot exams (59%) in comparison to Whites (71%) and Blacks (77%).
    - In the same year, smaller disparities were seen according to age. Nearly 75% of all adults with diabetes between ages 65-74 received a foot exam, about 73% of adults between ages 45-64 and 71.5% of adults over age 75.
    - There were not significant disparities by gender: In 2010, 72.3% of males and 70.7% of females received foot exams.
    - Adults with an education greater than high school received foot exams at 70% while adults with only a

high school education received foot exams at 67.8%; this gap widens for adults that achieved less than a high school education with only 59.1% receiving foot exams.

- NCQA's Diabetes Recognition Program 2015, 2016, and 2017. The mean ranged from 71.7% to 75.2% • Developer provided performance data also from the 2015 PQRS reporting year with a mean of 56.3%. Disparities: • In 2010, Hispanics had the lowest percentage of foot exams (59%) in comparison to Whites (71%) and Blacks (77%). o In the same year, smaller disparities were seen according to age.Education greater than high school received foot exams at 70% while adults with only a high school education received foot exams at 67.8%; this gap widens for adults that achieved less than a high school education with only 59.1% receiving
- The current measure performance is inconsistent across years.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- Performance data from NCQA DRP 2015 - 2017 and PQRS 2015
- Evidence of educational (reflection of socioeconomic) disparities and age disparities and race disparities
- Moderate rating
- Data from the NCQA's elite Diabetes Recognition Practices demonstrated mean completion of all three components of the measure between 71.7% and 75.2% in 2015, through 2017. Given that these reporting practices have achieved designation, their completion rate is likely much higher than the national mean.
- PQRS reporting (2015) demonstrated a mean completion of all 3 elements of 56.3 %
- There is substantial room for improvement
- Disparities data was provided from CDC stats, but the source of the data was not identified by the developer. Was this from the BRFSS?
- Performance data from NCQAs Diabetes Recognition Program from 2015-2017 are provided. PQRS data from its final year of 2016 are also provided. It is not clear if the data from these sources can be generalized to other clinical settings and provider populations. Comprehensive diabetes care continues to be a HEDIS measure in 2018. CDC data from 2010 is cited that compares the percentages of adult diabetic patients receiving a foot exam in a given year by age, sex, education levels, and race/ethnicity.
- Yes, there appears to be both suboptimal performance as a whole, as well as disparities between subpopulations reflecting disparities in care.
- Performance in the NCQ a diabetes recognition program showed me and it ranges of 71 to 75%. PQRS performance was less at 56.3%. Age, education and ethnicity were found to be areas of disparity.

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.

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

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: Primary Care and Chronic Illness project team staff

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

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

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

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

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

Committee pre-evaluation comments

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

Reliability Specs

- Data elements are clearly defined. No concerns with the codes used to identify the outcomes. No concerns with the ability to consistently implemented though it is dependent on chart data and how well foot exams are documented in the chart.
- No concerns regarding specifications
- Concur with the analysis of the staff evaluator.
- No concerns; empirical reliability testing done with NCQA DRP
- The measure components are easy to complete during an office visit and easy to capture with specific EHR fields. If the EHR lacks specific fields, the data is subject to poor reliability in its reporting. DRP practices may have this reporting mechanism as a priority in their HER but many practices will not.
- Reliability of data from the NCQA Diabetes Recognition Program is explained.
- None
- Essentially unchanged from previous review. No concerns.

Reliablity Testing

- I have no concerns related to the reliability of this measure.
- No concerns with the reliability of the measure.
- Concur with the analysis of the staff evaluator.
- no concerns
- The high reliability of DRP practices of .91 demonstrates real differences, but that is not necessarily generalizable to non DRP practices
- No
- Essentially unchanged from previous review. No concerns.

Validity Testing

- I have no concerns related to the validity of this measure
- Missing data does not constitute a threat to the validity of the measure.
- No concerns regarding validity testing
- Concur with the analysis of the staff evaluator.
- no concerns
- Construct and face validity tested
- didn't specifically test for distortion by exclusions (frequency of occurrence, variability of exclusions by provider, sensitivity analysis with and without exclusions) however, face validity suggests the exclusions are supported by the evidence
- The results indicate that the multiple experts, stakeholders and NCQA's Clinical Programs Committee concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.
- Construct validity was demonstrated along with medical nephropathy and eye exam.
- Validity of data from the NCQA Diabetes Recognition Program is explained.
- Essentially unchanged from previous review. No concerns.

Other threats

- I have no concerns related to the validity of this measure.
- Not applicable.
- No threats to validity from exclusions or risk adjustment
- Why is there a two outpatient visit versus one outpatient visit in the one data source? It would seem if it is this important that one visit would warrant the need for the foot exam.
- NA
- Exclusions are appropriate.
- There was no analysis in light of Social risk factors. The data extracted from DRPs demonstrated a .91 high degree of reliability.
- No risk adjustment was discussed

Criterion 3. feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

Data Specifications and Elements

- The measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records)
- Some data elements are in defined fields in electronic sources
- Developer shared no difficulties on the use of this measure in CMS QPP or NCQA's Diabetes Recognition Program.
- This is not an eMeasure

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- The required data elements are generated and used in the routine delivery of care.

- The data elements related to this process measure are generated as a byproduct of care delivery.
- The measure is feasible and it has been reported for years. The challenge is the resource use and cost involved in collecting the data as it is uses both chart and claims data.
- data elements are routinely generated and with the more widespread use of electronic health records, data collection should be more efficient
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- No concerns; developer can readily access data via CMA QPP and NCQA DRP;
- Moderate rating
- Feasibility of performance is very high.
- Feasibility of data collection and reporting is largely dependent on the presence of fields in the EHR to capture performance. As the developer comments, its presence is likely to increase over time.
- To allow for widespread reporting across physicians and clinical practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records).
- Since looks at an exam finding, this is often captured in free text and can be challenging to get in form of structured data without requiring extra work for the clinician. Most EMRs have a workaround for finding ways to more easily document this information.
- No concerns, unless foot exam is not a common data element within the health record being used by the providers.

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
CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

This measure is also used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes.

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
This measure uses the following methods to obtain input: including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.

Questions received through NCQA's Policy Clarification Support system have generally centered around clarification on what constitutes a foot exam, whether documentation must specify that all three exams (visual inspection and sensory exam with mono filament and a pulse exam) were completed, and if a mono filament is required for the sensory exam. In response, the developer has provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

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

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

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

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

4b. Usability 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 developer states that performance rates have stayed stable, despite a decrease in the number of reporting physicians seeking recognition in the NCQA's Diabetes Recognition Program since 2015.

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
- Per developer, there were no identified unexpected findings (positive or negative) during testing or since implementation of this measure.

Potential harms
- The developer did not identify any potential harms in testing.

Additional Feedback:
NA

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- This measure is currently used in public reporting and accountability programs.
- This measure is used on public reporting and accountability programs.
- In use for the Diabetes Recognition Program and the CMS Quality Payment Program. Used in HEDIS reporting. HEDIS reporting allows for health plans to provide feedback on measure results.
- This measure is widely reported and performance results are available for review. Feedback is provided by NCQA
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity
- already in use CMS QPP and NCQA DRP
- Utilizes NCQA Policy Clarification Support system;
- NCQA uses advisory panels, public commenting, several multistakeholder groups
- Publicly reported via CMS QPP and NCQA DRP portal.
- Multi stakeholder advisory panels and public commenting periods."
- The Diabetes Recognition Program has more than 10,000 clinician members. There are DRP publications and monthly webinars.
- Available for review by providers, as well as publicly reported data.

Usability

- There are no unintended consequences related to this measure.
- There are no evident unintended consequences related to this measure.
- In use for the Diabetes Recognition Program and the CMS Quality Payment Program. No unintended consequences sited by the Developer.
- There is a direct relationship between diabetic foot problems due to circulation or neuropathic changes and risk for amputation. Performance scores need improvement. No harms result from this measure.
- There is no discussion of challenges to improvement – Measure Developer reports that performance is stable which does not mean improved.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- Results provide feedback to provider of quality of diabetic foot care delivered
- No potential harms
- Rates of performance have remained stable although fewer practices are seeking recognition as a DRP. This will, over time require more dependence on EHR recording and reporting.
- There are no harms or unintended consequences.
- Since 2015, there has been a decrease in the number of reporting physicians seeking recognition in the DRP but rates in performance have remained relatively stable. There were no identified unexpected findings during testing or since implementation of this measure.
- There is room for performance in this measure, no unanticipated harms.

Criterion 5: Related and Competing Measures

Related or competing measures

- Developer identified one relating measure-0417 : Diabetic Foot & Ankle Care, Peripheral Neuropathy – Neurological Evaluation

Harmonization

- The developer noted difference between 0056 and 0417 in that measure 0056 identifies adults with diabetes (age 18-75) that had a foot exam (visual inspection with sensory and pulse exam) during the reporting year. Measure 0417 identifies adults with diabetes (age 18 and older) who had a lower extremity neurological exam at

least once during the measurement year.

- In addition data sources vary for these two measures. Measure 0056 is specified for paper medical records, administrative claims and electronic clinical data while measure 0417 is specified for administrative claims only.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0056
Measure Title: Comprehensive Diabetes Care-Foot Exam

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

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

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

   ☒Yes (go to Question #2)

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

   The measure's reliability per beta-binomial model is 0.91. This result indicates the measure has high reliability

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

Beta-binomial calculation was used to test measure score reliability.

3. Was reliability testing conducted with computed performance measure scores for each measured entity?
REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

Reliability was assessed from physician/practice data from the NCQA Diabetes Recognition Program that included 2866 physicians for the time frame of 2010-2012.

4. 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.*
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

The measure's reliability per beta-binomial model is 0.91. This result indicates the measure has high reliability, meaning that differences in physicin performance reflect true differences in quality as opposed to measurement error or noise.

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.
   ☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:
    ☒High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)
    ☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)
    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete]
    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

There was no testing of the exclusions done.

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

REFERENCE: Testing attachment, section 2b2.

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

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

There was no testing of the exclusions done.

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

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

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)
No missing data and "measure is collected with a complete sample" per developer.

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)
Developer did construct validity testing.

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

The developer tested for construct validity by exploring whether the measure was correlated with other similar measures of quality in NCQA's Diabetes Recognition Program hypothesized to be related.

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*

REFERENCE: *Testing attachment, section 2b1.*

TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

REFERENCE: Testing attachment, section 2c

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

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0056

Measure Title: Diabetes: Foot Exam

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

Date of Submission: 4/9/2018

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: Diabetic Foot Exam

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Adults with diabetes (type 1 or 2) >>> foot exam (visual inspection with sensory and pulse exam)>>> Exam results are evaluated >>>Results indicative of improper foot care >>>Health provider determines treatment to prevent further damage to the foot, such as possible infections or amputations>>>improvement in diabetes complications and quality of life.

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.)

N/A

****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.

N/A

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

Table 1. American Diabetes Association (ADA) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Diabetes Association. (2018). Standards of Medical Care in Diabetes – 2018. Diabetes Care 2018; 41(Suppl. 1): S105-S118; doi: 10.2337/dc18-S010 Guideline available from: http://care.diabetesjournals.org/content/41/Supplement_1 2013 Submission American Diabetes Association. (2013). Standards of Medical Care in Diabetes – 2013. Diabetes Care 2013; 36:S1-e4; doi: 10.2337/dc13-S001 Guideline available from: http://care.diabetesjournals.org/content/36/Supplement_1/S11
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.	2018 Submission • Perform a comprehensive foot evaluation at least annually to identify risk factors for ulcers and amputations. B • All patients with diabetes should have their feet inspected at every visit. C • Obtain a prior history of ulceration,

	amputation, Charcot foot, angioplasty or vascular surgery, cigarette smoking, retinopathy, and renal disease and assess current symptoms of neuropathy (pain, burning, numbness) and vascular disease (leg fatigue, claudication). B • The examination should include inspection of the skin, assessment of foot deformities, neurological assessment (10-g monofilament testing with at least one other assessment: pinprick, temperature, vibration), and vascular assessment including pulses in the legs and feet. B • Patients with symptoms of claudication or decreased or absent pedal pulses should be referred for anklebrachial index and for further vascular assessment as appropriate. C • A multidisciplinary approach is recommended for individuals with foot ulcers and high-risk feet (e.g., dialysis patients and those with Charcot foot, prior ulcers, or amputation). B • Refer patients who smoke or who have histories of prior lower-extremity complications, loss of protective sensation, structural abnormalities, or peripheral arterial disease to foot care specialists for ongoing preventive care and life-long surveillance. C • Provide general preventive foot self-care education to all patients with diabetes. B • The use of specialized therapeutic footwear is recommended for highrisk patients with diabetes including those with severe neuropathy, foot deformities, or history of amputation. B 2013 Submission Pg S8-S9 • "For all patients with diabetes, perform an annual comprehensive foot examination to identify risk factors predictive of ulcers and amputations. The foot examination should include inspection, assessment of foot pulses, and testing for loss of protective sensation (LOPS) (10-g monofilament plus testing any one of the following: vibration using 128-Hz tuning fork, pinprick sensation, ankle reflexes, or

	vibration perception threshold). (B) • Provide general foot self-care education to all patients with diabetes. (B) • A multidisciplinary approach is recommended for individuals with foot ulcers and high-risk feet, especially those with a history of prior ulcer or amputation. (B) • Refer patients who smoke, have LOPS and structural abnormalities, or have a history of prior lower-extremity complications to foot care specialists for ongoing preventive care and lifelong surveillance. (C) • Initial screening for peripheral arterial disease (PAD) should include a history for claudication and an assessment of the pedal pulses. Consider obtaining an ankle-brachial index (ABI), as many patients with PAD are asymptomatic. (C) • Refer patients with significant claudication or a positive ABI for further vascular assessment and consider exercise, medications, and surgical options. (C)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Level of evidence and description: • B: Supportive evidence from well-conducted cohort studies, including: o Evidence from a well-conducted prospective cohort study or registry o Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study • C Supportive evidence from poorly controlled or uncontrolled studies o Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results o Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) o Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation

	2013 Submission Same as above
Provide all other grades and definitions from the evidence grading system	2018 Submission Level of Evidence & Description: • A: Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted multicenter trial o Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted trial at one or more institutions o Evidence from a meta-analysis that incorporated quality ratings in the analysis • E: Expert consensus or clinical experience 2013 Submission Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission

	No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The ADA does not provide information on the systematic review conducted to support its 2018 or 2013 guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on one other systematic review that support the ADA's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 3 below
What harms were identified?	See Table 3 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 2. American Geriatrics Society (AGS)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Geriatrics Society (AGS). 2013. Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. American Geriatrics Society Panel on the Care for Older Adults with Diabetes Mellitus. Journal of American Geriatric Society. 2013 November; 61 (11): 2020-2026. Doi:10.1111/jgs.12514 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064258/pdf/nihms583558.pdf 2013 Submission American Geriatrics Society (AGS). 2003. Guidelines for Improving the Care of the Older Person with Diabetes Mellitus. California Healthcare Foundation/American Geriatrics Society Panel on Improving Care for Elders with Diabetes. American Geriatrics Society. May 2003; 51, Suppl 5, JAGS URL
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions	2018 Submission • Older adults with DM should have a careful foot examination at least annually to check skin integrity and to determine whether there is loss of sensation

from the SR.	or decreased perfusion and more frequently if there is evidence of any of these findings (IIIA) 2013 Submission Pg S272 • "The older adult who has DM should have a careful foot examination at least annually to check skin integrity and to determine whether there is bony deformity, loss of sensation, or decreased perfusion and more frequently if there is evidence of any of these findings. (IIIA) There are no RCT data to support examination of the feet at regular intervals to prevent lower-extremity ulceration or amputation, but a randomized trial of an intervention consisting of patient and provider foot-care education and a team approach to foot care found an increase in rates of foot examinations at routine office visits and a reduction in serious foot lesions (odds ratio (OR) = 0.41, *P* =.05). In addition, several uncontrolled studies have found a reduction in rates of amputation after implementation of comprehensive foot-care programs. Regular foot examinations permit identification of diabetic neuropathy and foot lesions and may in turn prevent progression to ulcers and amputation, but there are no data to support the optimal interval for evaluation. Most current recommendations specify that the foot examination should be done at all nonurgent outpatient visits. Quality of evidence is level II for more frequent examinations for persons at high risk for foot problems and level III for routine annual screening."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Quality of Evidence • Level III: Evidence from respected authorities based on clinical experience, descriptive studies, or reports of expert committees Strength of Evidence • A: Good evidence to support the use of a recommendation; clinicians should do this all the time 2013 Submission Same as above

Provide all other grades and definitions from the evidence grading system	2018 Submission Quality of Evidence • Level I: Evidence from at least one properly randomized controlled trial • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytical studies, from multiple time-series, or from dramatic results in uncontrolled experiments Strength of Evidence • B: Moderate evidence to support the use of a recommendation clinicians "should do this most of the time" • C: Poor evidence to support or to reject the use of a recommendation; clinicians may or may not follow the recommendation • D: Moderate evidence against the use of a recommendation; clinicians should not do this • E: Good evidence against the use of a recommendation; clinicians should not do this 2013 Submission Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AGS does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the AGS systematic review, we provide information on two other systematic reviews that support the AGS's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 3 below
What harms were identified?	See Table 3 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 3. Additional Systematic Reviews

Citations	Singh, N., Armstrong, D. G., & Lipsky, B. A. (2005). Preventing foot ulcers in patients with diabetes. *JAMA: the journal of the American Medical Association*,*293*(2), 217-228.
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	This systematic evidence review focused on the efficacy of methods advocated for preventing diabetic foot ulcers. We will present the evidence for the efficacy of screening to identify patients at risk for diabetic foot ulcers and two specific clinical interventions to prevent foot ulcers (foot examination by a clinician and foot specialist/multidisciplinary care team). Additional information (not presented here) can be found in the review on the effectiveness of optimizing glycemic control, smoking cessation, custom footwear, debridement of calluses, and surgery on reducing the incidence of foot ulcers.
Grade assigned for the quality of the quoted evidence with definition of the grade	No grading provided.
Provide all other grades and associated definitions of the evidence in the grading system	N/A
What is the time period	1980-2004

<table>
<tr><td>covered by the body of evidence?</td><td></td></tr>
<tr><td>Quantity and Quality of Body of Evidence</td><td>Studies related to efficacy of screening to identify patients at risk for diabetic foot ulcer: 5 prospective cohort studies; 2 case control studies

Studies related to clinical interventions to prevent food ulceration: 3 RCTs; 1 case-control; 1 cohort study</td></tr>
<tr><td>What is the overall quality of evidence <u>across studies</u> in the body of evidence?</td><td>The authors of the review did not comment on the quality of the evidence related to efficacy of screening to identify patients at risk for foot ulcers. However, authors concluded the evidence from the seven studies of screening test efficacy was strong enough to support the use of screening to identify patients at risk.

The case-control study of the effectiveness of foot examination by a clinician did not show any significant reduction in amputation among 244 diabetic patients (OR 0.55; 95% CI, 0.2-1.7; P=.31). However, the study was limited by high rates of foot examination in both case and control patients, different degree of risk between the groups as well as the unusually high rates of diabetes and amputation among the Pima Indian population included in the study.

The three RCTs of clinician and specialist intervention were of reasonable size (N=91-498) and good quality.</td></tr>
<tr><td>Estimates of benefit and consistency across studies in body of evidence – what are the estimates of benefits?</td><td>Evidence related to efficacy of screening to identify patients at risk for diabetic foot ulcer: The authors summarized the efficacy of different screening methods in the table below. They conclused that the monofilament test is the most validated test, however the number of test sites needed for the test is still unclear. The Biothesiometer test has similar accuracy to the monofilament test, but is not as widely available. The Tuning form and pressure mat tests are not as accurate.

<table>
<tr><td>Screening Method to Identify Patients at Risk for Diabetic Foot Ulcer</td><td>Monofilament (Light Touch Sensation)</td><td>Biothesiometer (Vibratory Sensation)</td><td>Tuning Fork (Vibratory Sensation)</td><td>Pressure Mat or Platform (Plantar Pressure)</td></tr>
<tr><td>Sensitivi</td><td>66-91</td><td>83-86</td><td>55-61</td><td>57-64</td></tr>
</table>
</td></tr>
</table>

	ty, %				
	Specificity, %	34-86	57-63	59-72	46-70
	Positive Predictive Value %	18-39	20-32	16	17-31
	Negative Predictive Value %	94-95	95-97	93	82-90

Evidence Related to Clinical interventions to prevent food ulceration: One randomized study of diabetic persons (N=91) with a previous foot ulceration found a significantly reduced risk for ulceration recurrence (RR, 0.52; 95% CI, 0.29-0.93; P = .03) at 1 year for those who received routine podiatric care. In another randomized study trial of diabetic persons with neuropathy (N=498), patients randomized to receive podiatric care at least twice a year had no difference in the incidence of foot ulcers compared to usual care, but fewer deep ulcers (6 vs 12), infected ulcers (1 vs 10; P_.01), and hospital admission days (24 vs 346; P_.01) compared to usual care patients. A cohort study included diabetic persons (N=341) who were examined to categorize baseline risk, initiate appropriate education and interventions, and schedule follow-up foot examinations and podiatric care with a multidisciplinary team. After 3 years, the incidence of lower-extremity amputation was only 1.1 per 1000 persons per year. Among high-risk persons, those who missed more than 50% of their appointments with the team were 54 times more likely to develop an ulcer and 20 times more likely to require an amputation than those who kept most appointments.

What harms were studied and how do they affect the net benefit (benefits over harms)?	There were no harms to screening or clinician examination reported in the review.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous studies have been conducted since the systematic review we cite in this table, none of which change the conclusion that routine foot exams for individuals with diabetes are appropriate. Below we list two additional studies that support this measure. Sloan FA, Feinglos MN, Grossman DS. Receipt of care and reduction of lower extremity amputations in a nationally representative sample of U.S. Elderly. Health Serv Res. 2010;45(6 pt 1):1740-1762.

	Carls GS, Gibson TB, Driver VR, et al. The economic value of specialized lower-extremity medical care by podiatric physicians in the treatment of diabetic foot ulcers. J Am Podiatr Med Assoc. 2011;101:93-115.

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.

N/A

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

N/A

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

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

N/A

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0056
Corresponding Measures:
De.2. **Measure Title:** Diabetes: Foot Exam
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a foot exam (visual inspection and sensory exam with mono filament and a pulse exam) during the measurement year.
1b.1. **Developer Rationale:** This measure promotes regular foot examinations in adults with diabetes (ages 18-75). Because of macrovascular compromise leading to arterial insufficiency and microvascular effects on nerve function, surveillance of skin integrity is very important for patients with diabetes. Poor foot care can lead to infections and ultimately amputations of the toe, foot, lower limb, or upper limb. As a result of amputations, patients often experience drastic declines in quality of life. In order to maintain optimal quality of life for persons with diabetes, it is vital to maintain the highest quality of foot care in diabetic populations.

S.4. **Numerator Statement:** Patients who received a foot exam (visual inspection and sensory exam with monofilament and pulse exam) during the measurement period.
S.6. **Denominator Statement:** Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year.
S.8. **Denominator Exclusions:** -Patients with a diagnosis of secondary diabetes due to another condition (e.g. a diagnosis of gestational or steroid-induced diabetes)
-Patients who have had either a bilateral amputation above or below the knee, or both a left and right amputation above or below the knee before or during the measurement period.
-Exclude patients who were in hospice care during the measurement year

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data, Other, Paper Medical Records
S.20. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

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?**

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_0056_Foot_Exam_7.1.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.
No

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 promotes regular foot examinations in adults with diabetes (ages 18-75). Because of macrovascular compromise leading to arterial insufficiency and microvascular effects on nerve function, surveillance of skin integrity is very important for patients with diabetes. Poor foot care can lead to infections and ultimately amputations of the toe, foot, lower limb, or upper limb. As a result of amputations, patients often experience drastic declines in quality of life. In order to maintain optimal quality of life for persons with diabetes, it is vital to maintain the highest quality of foot care in diabetic populations.

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 is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care (see full description of program in 4a1.1). Below is performance data for this measure in the program.

Diabetes Recognition Program
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2015|4989|74.3%|28.8%|0.0%|22.9%|65.4 |84.6%|95.4%|100.0%|100.0%
2016|4458|71.7%|29.2%|0.0%|20.0%| 56.0%|84.0%|92.8%|98.9%|100.0%
2017|3971|75.2%|25.9%|0.0%|32.0%|64.0 |84.0%|94.8%|100.0%|100.0%

PQRS
The following PQRS performance data includes claims, registry, measures group, GPRO Web Interface/ACO, QCDR data for services performed from in 2015.
Mean: 56.3%
St dev: 32.0%

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.**
N/A

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.*
While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if those data are available to a practice. See response in 1b.5 for more information.

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 Centers for Disease Control and Prevention examined the proportion of diabetic adults (over age 18) that received a foot exam in a given year. This data was categorized based on race/ethnicity, age, sex, and education level. In 2010, Hispanics had the lowest percentage of foot exams (59%) in comparison to Whites (71%) and Blacks (77%) (CDC, 2012). In the same year, smaller disparities were seen according to age. Nearly 75% of all adults with diabetes between ages 65-74 received a foot exam, about 73% of adults between ages 45-64 and 71.5% of adults over age 75 (CDC, 2012). There were not significant disparities by gender: In 2010, 72.3% of males and 70.7% of females received foot exams (CDC, 2012). Adults with an education greater than high school received foot exams at 70% while adults with only a high school education received foot exams at 67.8%; this gap widens for adults that achieved less than a high school education with only 59.1% receiving foot exams (CDC, 2012).

Centers for Disease Control and Prevention (CDC). 2012. CDC's Diabetes Program-Data and Trends-Prevalence of Diabetes-Percent of Foot Exam in the Last Year for Adults Aged =18 Years, by Race/Ethnicity. Retrieved from http://www.cdc.gov/diabetes/statistics/preventive/tNewFtChkRace.htm.

Centers for Disease Control and Prevention (CDC). 2012. CDC's Diabetes Program-Data and Trends-Prevalence of Diabetes-Percent of Foot Exam in the Last Year for Adults Aged =18 Years, by Age. Retrieved from http://www.cdc.gov/diabetes/statistics/preventive/tNewFtChkAgeTot.htm .

Centers for Disease Control and Prevention (CDC). 2012. CDC's Diabetes Program-Data and Trends-Prevalence of Diabetes-Percent of Foot Exam in the Last Year for Adults Aged =18 Years, by Sex. Retrieved from http://www.cdc.gov/diabetes/statistics/preventive/tNewFtChkSex.htm .

Centers for Disease Control and Prevention (CDC). 2012. CDC's Diabetes Program-Data and Trends-Prevalence of Diabetes-Percent of Foot Exam in the Last Year for Adults Aged =18 Years, by Education. Retrieved from http://www.cdc.gov/diabetes/statistics/preventive/tNewFtChkEduc.htm.

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):*
Endocrine, Endocrine : Diabetes

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):*
Populations at Risk

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.)*

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0056_CDC_Foot_Exam_Value_Set_.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

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

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who received a foot exam (visual inspection and sensory exam with monofilament and pulse exam) during 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)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Time period for data: a measurement year (12 months)
ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying numerator events for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD: At a minimum, documentation in the medical record must include a note indicating the date when the exam was performed and the result. The patient is numerator compliant if a foot exam during the measurement year and result are documented. The patient is not numerator compliant if the result for the foot exam and result during the measurement year are missing. Ranges and thresholds do not meet criteria for this measure.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
ENCOUNTER: Patients who had a visit (office visit, face to face encounter, preventive care services, home healthcare services, annual wellness) during the measurement period

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES:
Alpha-glucosidase inhibitors:
Acarbose, Miglitol

Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metaformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metaformin-saxagliptin, Metformin-sitagliptin , Sitagliptin-simvastatin

Insulin:
Insulin aspart, Insulin aspart-insulin aspart protamine, insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitagliptin

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
-Patients with a diagnosis of secondary diabetes due to another condition (e.g. a diagnosis of gestational or steroid-induced diabetes)
-Patients who have had either a bilateral amputation above or below the knee, or both a left and right amputation above or below the knee before or during the measurement period.
-Exclude patients who were in hospice care during the measurement year

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD

Exclusionary evidence in the medical record must include a note indicating a diagnosis of gestational or steroid-induced diabetes, patients who had either a bilateral amputation above or below the knee, or both a left and right amputation above or below the knee, or who are in hospice care.

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

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 = Higher 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.)*
STEP 1. Determine the eligible population. To do so, identify patients who meet all the specified criteria.
-AGES: 18-75 years as of December 31 of the reporting period.
-EVENT/DIAGNOSIS:
Identify patients who had a diagnosis of diabetes with a visit during the measurement period.
*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S2.B
STEP 2. Determine the number of patients in the eligible population who had a recent foot exam (visual inspection with a sensory exam and a pulse exam) exam during the measurement year through the search of administrative data systems.
STEP 3. Identify patients with a most recent foot exam performed and the result.
STEP 4. Identify the most recent foot exam with a result during the reporting period (numerator compliant). Identify the most recent result foot exam without a result or a missing foot exam (not numerator compliant).
STEP 5. Exclude from the eligible population patients from step 2 for whom administrative system data identified an exclusion to the service/procedure being measured. *SEE DENOMINATOR EXCLUSION CRITERIA IN QUESTION S.9
STEP 6. Calculate the rate (number of patients that received a foot exam during the measurement year).

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

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 Data, Paper Medical 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.
This measure uses a combination of administrative claims data and medical records. Foot exams can be identified by the following administrative data: receipt of a foot exam (visual inspection and sensory exam with mono filament and a pulse exam).

Codes in the following value set will meet these criteria:
-Any code in the Physical Exam, Performed: Visual Exam of Foot value set
-Any code in the Physical Exam, Performed: Sensory Exam of Foot
-Any code in Physical Exam, Performed: Pulse Exam of Foot

The minimum medical record documentation includes a note indicating the date when the exam was performed and the result.

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

No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, 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.)*
N/A

2. Validity – See attached Measure Testing Submission Form
0056_Foot_Exam_2018_Testing_Form.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0056
Measure Title: Comprehensive Diabetes Care: Foot Exam
Date of Submission: 3/5/2018
Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ 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? 2010-2012

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*)

We calculated measure score reliability from physician/practice level data from the NCQA Diabetes Recognition Program (DRP) that included 2866 physicians. Construct validity was calculated with data from a sample of 653 physicians/practices.

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

NCQA's Diabetes Recognition Program currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes. Below is a description of the sample. It includes the number of physicians and practices reporting on this measure in the DRP program in 2012.

Analysis	Number of physicians	Median denominator size
Reliability	2,866	25
Construct Validity	653	25

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

Reliability was tested using a beta-binomial calculation. This analysis included the entire DRP sample described above.

Validity was demonstrated through construct validity using data from a sample of 653 physicians/practices and through a systematic assessment of face validity with expert panels.

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

We did not analyze performance by social risk factors.

2a2. RELIABILITY TESTING

***<u>Note</u>**: 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*)

Reliability was estimated by using the beta-binomial model for the physician/practice level Diabetes Recognition Program measure. The beta-binomial model assumes the performance score is a binomial random variable conditional on the true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped. This is the general way we look at beta-binomial data, which is typically drawn from a group with varied performance. We will note however, that for this measure, our data does not comply with those assumptions.

Reliability as we've described it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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*)

Product Type	**Reliability per Beta Binomial Model**
Diabetes Recognition Program	0.91

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 value for the beta-binomial statistic for the physician level measure suggests the measure has high reliability.

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)

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality in NCQA's Diabetes Recognition Program hypothesized to be related, which are listed below.

- Eye Exam
- Medical Attention for Nephropathy

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of Assessing Face Validity

This measure was tested for face validity with four panels of experts. The Diabetes Recognition Program (DRP) Advisory Committee included 7 experts in diabetes care including representation by clinicians, health plans, integrated health systems and research organizations; Diabetes Measurement Advisory Panel (DMAP), Committee on Performance Measurement (CPM) and the Clinical Programs Committee (CPC). All measures incorporated in NCQA programs benefit from a 30-day public comment period and real-time feedback from our Policy Clarification System portal that receives over 3,500 inquiries annually. NCQA's CPC's oversees the evolution of NCQA's recognition programs and related measures including the Diabetes Recognition Program, the Heart/Stroke Recognition Program, the Patient Centered Medical Home and Patient-Centered Specialty Practice Recognition Program, among others. The CPC includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 18 members. The CPC is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of clinical recognition programs. CPC members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel

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

Construct Validity

Table 1 below provides the results from construct validity testing of the physician level measure.

Table 1. Correlations among Diabetes Measures in the NCQA Diabetes Recognition Program - 2012

	Pearson Correlation Coefficient	
	Eye Exam	Medical Attention to Nephropathy
CDC – Foot Exam	0.42	0.29

Note: All correlations are significant at p<0.0001

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?*)

Construct Validity
The *CDC-Foot Exam* measure has a moderate correlation with the *Eye Exam* measures in the Diabetes Recognition Program. The correlation between the *Foot Exam* measure and the *Medical Attention to Nephropathy* measure is just under the .3 value and indicates a slightly weaker but still relevant association. Overall, these correlation results suggest that the physician level measure has sufficient validity.

Face Validity
The results indicate that the multiple experts, stakeholders and NCQA's Clinical Programs Committee concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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*)

Testing was not performed for the excluded sample.

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*)

Testing was not performed for the excluded sample.

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*)

Testing was not performed for the excluded sample.

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

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

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**
☐ **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*)
N/A

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample, size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans performance is significantly different from each other.

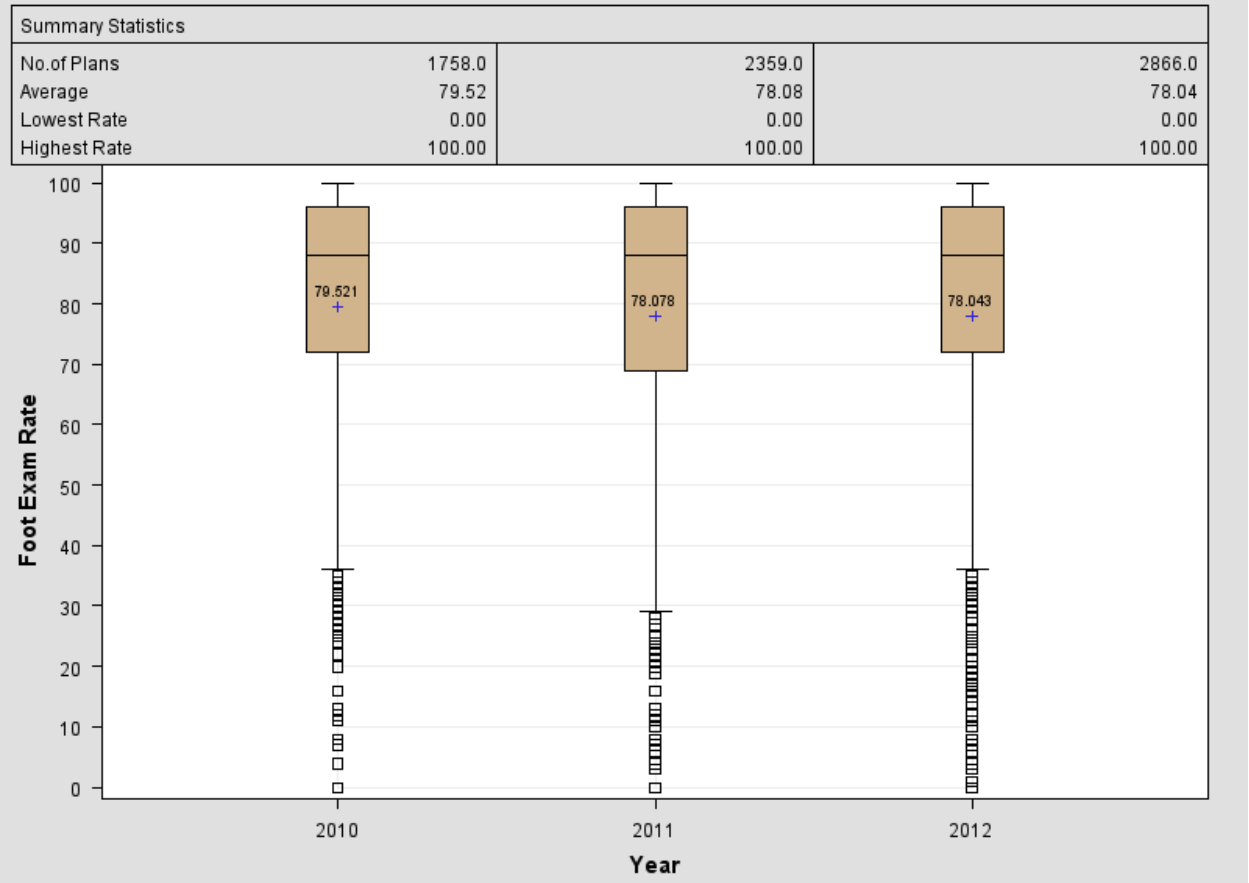
2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly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

Year	N (# of clinicians	Mean (%)	St Dev (%)	Min (%)	10th (%)	25th (%)	50th (%)	75th (%)	90th (%)	Max (%)	IQR (%)	p value
2010	1763	79.52	23.61	0.0	43.0	72.0	88.0	96.0	100.0	100.0	24.0	<0.05
2011	2359	78.08	25.39	0.0	36.0	69.0	88.0	96.0	100.0	100.0	27.0	<0.05
2012	2866	78.04	25.56	0.0	36.0	72.0	88.0	96.0	100.0	100.0	24.0	<0.05

IQR: Interquartile range

p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

Chart 1. Boxplot of Foot Exam Measure, Diabetes Recognition Program, 2010-2012



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 difference between the 25th and 75th percentile is statistically significant, suggesting there are meaningful differences in performance.

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

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*)

This measure is collected with a complete sample.

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*)

This measure is collected with a complete sample.

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*)

This measure is collected with a complete sample.

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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**.

Some data elements are in defined fields in electronic sources

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).

To allow for widespread reporting across physicians and clinical practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

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

Attachment:

3c. Data Collection Strategy

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

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

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

Feedback on use of this measure in CMS QPP and NCQA's Diabetes Recognition Program has been positive with few questions raised by participating clinicians to the CMS vendor and NCQA. NCQA also works with the CMS vendor to review any questions or issues raised with the measure on a bi-weekly basis.

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)*.

N/A

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 CMS Quality Payment Program https://qpp.cms.gov/ CMS Quality Payment Program https://qpp.cms.gov/ Professional Certification or Recognition Program Diabetes Recognition Program http://www.ncqa.org/Programs/Recognition/DiabetesRecognitionProgramDRP.aspx

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

CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

DIABETES RECOGNITION PROGRAM: This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care. The program currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. The DRP Program has 6 measures which cover other areas such as: HbA1c control, blood Pressure control, eye examinations, nephropathy assessment, smoking and tobacco use and cessation advice or treatment, and foot examinations. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes.

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?*)
N/A

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.*)
N/A

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
The Diabetes Recognition Program (DRP) data submission portal provides a guided process for practice/clinician to submit their patient data (manually or electronically). Practices/Clinicians can see their results within the data portal. The portal is equipped to immediately score the data to determine if it meets the measure performance requirements. The DRP publication provides instruction on the required data points for this measure, reference to guidelines used to curate the measure requirements and additional information for achieving recognition. NCQA provides monthly webinars to instruct customers on the measure, the specifications, data entry in the portal and recognition readiness. These live webinars also provide the opportunity to ask additional questions related to DRP, the measure, the recognition process and other program related questions.

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.**
For the Diabetes Recognition Program (DRP) practices/clinicians submit data (manually or electronically) when they are ready and

their data meet the performance thresholds for the measure. Practices/clinicians can see their results within the data portal. The portal is equipped to immediately score the data to determine if it meets the measure performance requirements. Practices/Clinicians can see the results for this measure and remaining measures in DRP to determine if the entity has met the required score to achieve DRP recognition. Additional questions can be submitted for response using NCQA's Policy Clarification System (PCS). PCS providers NCQA staff and customers a unified space to submit inquires and clarification requests.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Questions received through NCQA's Policy Clarification Support system have generally centered around clarification on what constitutes a foot exam, whether documentation must specify that all three exams (visual inspection and sensory exam with mono filament and a pulse exam) were completed, and if a mono filament is required for the sensory exam.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the PQRS and the Diabetes Recognition Program.

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.**
We have provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

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.
Since 2015, there has been a decrease in the number of reporting physicians seeking recognition in NCQA's Diabetes Recognition Program (see summary data in 1b.2). However, we are pleased that rates in performance have remained relatively stable.

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.**
There were no identified unexpected findings during testing or since implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits during testing or since implementation of this measure.

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)**
0417 : Diabetic Foot & Ankle Care, Peripheral Neuropathy – Neurological Evaluation

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?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Measure 0056 identifies adults with diabetes (age 18-75) that had a foot exam (visual inspection with sensory and pulse exam) during the reporting year. Measure 0417 identifies adults with diabetes (age 18 and older) who had a lower extremity neurological exam at least once during the measurement year.

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.)
0056 has a long history of use and is implemented in two national programs (PRQS and DRP).

RESPONSE TO 5a.2 (insufficient space above)
Measure 0056 identifies adults with diabetes (age 18-75) that had a foot exam (visual inspection with sensory and pulse exam) during the reporting year. Measure 0417 identifies adults with diabetes (age 18 and older) who had a lower extremity neurological exam at least once during the measurement year.

HARMONIZED ELEMENTS:
Both measures are harmonized on the target population of diabetic adults and the measure focus of lower extremity exam. The denominator for each measure are harmonized to include all adult patients with a diagnosis of diabetes mellitus. The care setting is harmonized for measure 0056 and 0417 in at least one care setting (Ambulatory Care: Clinician Office/ Clinic). In addition, the data source (administrative claims) and level of analysis (clinicians: individual) are harmonized for both measures.

UNHARMONIZED MEASURE ELEMENTS:

Data Source: Measure 0056 is specified for paper medical records, administrative claims and electronic clinical data while measure 0417 is specified for administrative claims only. Measure 0056 is included in the CMS PQRS program and in NCQA's Diabetes Recognition Program (DRP) for physician reporting.

IMPACT ON INTERPRETABILITY AND DATA COLLECTION BURDEN: Measure 0056 provide more options for reporting based on available data sources. Measure 0417 is specified for only administrative claims.

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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
DIABETES EXPERT PANEL:
Bill Herman (Chair), MD, Univ. of Michigan Health System
David Aron, MD, Department of Veteran's Affairs
James Fain, PhD, RN, University of Massachussetts
Jerry Cavallerano, OD, Beetham Eye Institute
John Thompson, MD, Retina Specialists
Judith Fradkin, MD, NIDDK/NIH
Lynne Levitsky, MD, Massachusetts General Hospital
Mark Cziraky, PharmD, Healthcore
Richard Hellman, MD, Private Practice, Diabetes & Endocrinology
Seth Rubenstein, DPM, Reston Hospital Center, INOVA Fair Oaks Hospital
Stephen Fadem, MD, Baylor College of Medicine
Ted Ganiats, MD, Univ. of California, San Diego
Nancy Van Vessem, MD, Capital Health Plan

HEDIS EXPERT CODING PANEL
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Aqurate Health Data Management, Inc.
Nelly Leon-Chisen, RHIA, American Hospital Association
Alec McLure, MPH, RHIA, CCS-P, Verscend Technologies
Michele Mouradian, RN, BSN, Change HealthCare
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, WellCare Health Plans, Inc.

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers

Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

CLINICAL PROGRAMS COMMITTEE
Randall Curnow, MD, MBA, FACP, FACHE, FACPE (Chair), TriHealth
Suzanne Berman, MD, FAAP, Plateau Pediatrics
Brooks Daveman, MPP, Tennessee Division of Health Care Finance and Administration
Marcus Friedrich, MD, MBA, FACP, New York State Department Health Empire State Plaza, Coming Towne
Jennifer Gutzmore, MD, Cigna
Melissa Hogan, MPH, Aon
Adriana Matiz, MD, FAAP, Ambulatory Care Network
Lisa Morrise, Marts, LAM Professional Services, LLC
Deborah Murph, MBA, BSN, RN, Cherokee Health Systems
Amy Nguyen Howell, MD, MBA, CAPG
Marc Rivo, MD, Population Health Innovations
Julie Schilz, BSN, MBA, Anthem
Pamela Slaven-Lee, DNP, FNP-C, CHSE, The George Washington University School of Nnursing
Lina Walker, PhD, AARP

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 1999
Ad.3 **Month and Year of most recent revision:** 01, 2010
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2014

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested

for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** Publication of each Measure is to be accompanied by the following notice:

NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.
These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise modify the measures, and shall not disassemble, recompile or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0057
Corresponding Measures:
Measure Title: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Testing
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received an HbA1c test during the measurement year.
Developer Rationale: Testing hemoglobin A1c levels in patients with diabetes is an important component of diabetes treatment and care. Results from these tests aids clinicians in providing patients with optimal treatment that will maximize diabetes control and in turn prevent complications of diabetes that threaten to impact quality of life.

Numerator Statement: Patients who had an HbA1c test performed during the measurement year.
Denominator Statement: Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Members who do not have a diagnosis of diabetes in any setting, during the measurement year or the year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes in any setting, during the measurement year or the year prior to the measurement year.

Measure Type: Process
Data Source: Claims, Electronic Health Data, Paper Medical Records
Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- The developer provided updated guidelines from the American Diabetes Association (ADA) (2018) including recommendations for the following.
    - Perform the A1C test at least two times a year in patients who are meeting treatment goals (and who have stable glycemic control). **E**
    - Perform the A1C test quarterly in patients whose therapy has changed or who are not meeting glycemic goals. **E**
    - Point-of-care testing for A1C provides the opportunity for more timely treatment changes. **E**
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guideline, as this information was not available
    - Level of evidence -E (definition): Expert consensus or clinical experience
- The developer provided updated guidelines from 2013 from American Geriatrics Society
    - General Recommendations for "Glycemic Control"
        1. Target goal for glycosylated hemoglobin (HbA1c) in older adults generally should be 7.5% to 8%. HbA1c between 7% and 7.5% may be appropriate if it can be safely achieved in healthy older adults with few comorbidities and good functional status. Higher HbA1c targets (8–9%) are appropriate for older adults with multiple comorbidities, poor health, and limited life expectancy. (1A evidence for HbA1c 7–8%, and IIA for 8–9%) There is potential harm in lowering HbA1c to less than 6.5% in older adults with type 2 DM. **(IIA)**
        2. Older adults with DM whose individual targets are not being met should have their HbA1c levels measured at least every 6 months and more frequently as needed or indicated. For older adults with stable HbA1c over several years, measurement every 12 months may be appropriate. **(IIIB)**
        - Quality of Evidence
            - Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytic studies, or from multiple time-series studies, or from dramatic results in uncontrolled experiments
            - Level III: Evidence from respected authorities, based on clinical experience, descriptive studies, or reports of expert committee
        - Strength of Evidence
            - A: Good evidence to support the use of a recommendation; clinicians should do this all the time
            - B: Moderate evidence to support the use of a recommendation; clinicians should do this most of the time
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guideline, as this information was not available
- The evidence for this measure is based on measurement of blood glucose or HbA1c to facilitate glycemic control in adults with diabetes. Monitoring of blood glucose can be conducted by patients through self-monitoring (SBGM) or by the provider through point of care treatment (PoCT). Self-monitoring includes using at home blood glucose tests to continuously measure glucose levels. HbA1c tests are conducted or ordered by a provider to measure the average blood glucose over a three-month period. Results from monitoring assist providers and patients with maintaining or improving glycemic control and reducing complications from diabetes.
    - The VA/DoD evidence review gave this recommendation the following grading: **Level of Evidence=II, Quality of Evidence=fair, Strength of Recommendation=B**. The fair rating for the quality of evidence (see quality grading) indicates that the evidence can be linked to the health outcome. The B grading for this evidence signifies that HbA1c testing may be useful or effective. Furthermore, the level of evidence indicates that the studies used were well designed controlled trials, cohort or case controlled studies, or included multiple time series.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided updated guidelines from 2018 from the American Diabetes Association and from 2013 from the American Geriatrics Society which continues to support their measure focus.
- The developer also provided an additional systematic review from the Department of Veteran Affairs/Department of Defense (2010).

Exception to evidence

NA

Questions for the Committee:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*
  - *If derived from patient report, does the target population value the measured process or structure and find it meaningful?*

Guidance from the Evidence Algorithm

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

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

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

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

- Developer provided performance data extracted from HEDIS data, stratified by commercial health plan, Medicare, and Medicaid from 2014, 2015, and 2016.
  - Commercial performance
    - Mean: 89.42% (2014) to 89.91% (2016)
    - Standard Deviation: 3.73% (2014) to 3.53% (2016)
  - Medicaid performance
    - Mean: 86.31% (2014) to 86.66% (2016)
    - Standard Deviation: 4.84% (2014) to 5.66% (2016)
  - Medicare performance
    - Mean: 92.72% (2014) to 93.54% (2016)
    - Standard Deviation: 4.17% (2014) to 3.38% (2016)

Disparities

- Developer did not provide disparities data from the measure. However cited CDC data from 2000-2010 that reported percentage of diabetic adults over age 18 that received two or more hemoglobin A1c (HbA1c) tests within the last year by race/ethnicity, age, sex and education level.
    - The most recent data shows Hispanics with the lowest rates of HbA1c tests at 63.7%. Whites and Blacks had nearly equal percentages at 73.4% and 73.1%, respectively (CDC, 2012).
    - In 2010, adults in the 45-64 and 18-44 age groups had the lowest percentages for A1c tests at 72.4% and 63.5%, respectively. The highest percentage of A1c tests based on age was seen in the 65-74 age group (78.2%), followed by the over 75 age group (75.7%) (CDC, 2012).
    - Based on data from the CDC, women receive A1c tests at higher rates than men at 74% and 71.5% respectively (CDC 2012).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The updated evidence is stronger than it was for the previous review. With that in mind, there is no need for a repeat discussion and vote on the evidence.
- I am not aware of any additional evidence or studies other than what was submitted by the Developer.
- Evidence from the ADA 2018, AGS 2015 provided Moderately strong evidence regarding the importance of HgA1c testing. Control of blood sugar lessens and may prevent complication. The developer did not summarize the QQC of the body of evidence, as it was not provided in the clinical guidelines. The Evidence was rated "E" (ADA) expert consensus or clinical experience, IIIB (AGS) rating of evidence based on clinical experience, descriptive studies, reports of expert committee, B Moderate evidence to support use of reommendations clinicians should do most of the time. The VA/DOD evidene graded measuring HgA1c (or blood glucose Level II, Quality of evidence fair, strength of recommendation B. This last evidene review includes blood glucose self monitoring which is not the focus of this measure.
- What there any evidence regarding A1C Targets on Patient Experience? Some patients report that too tight a target ends in binge-like behavior that overall reduces quality of life and outcomes – a more personalized target based on patient experience and clinician recommendation may be more appropriate. With the AGS guideline being so specific about targeting – may need to consider what unintended consequences related to interpretation – i.e. setting too tight a target for a patient.
- No concerns on evidence; no need for repeat discussion and vote
- I agree that since the updated data is directionally the same, there is no need to vote on Evidence

Performance Gap

- Although overall performance is high on this measure, there does indeed appear to disparities by race/ethnicity. There also is disparity by age. With these disparities, it's important to continue performance measurement to serve as a basis to eliminate these disparities to understand why they exist.
- The Developer did not independently identify performance gaps other than between commercial, Medicare and Medicaid populations. They did provide CDC disparity gaps that included social factors.
- The CDC data suggests a gap in care/testing. I believe that the percentage difference great enough that there should be a national performance measure for Hispanics, men, adults between 18 and 64?"
- "Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?

- Education is not always a good proxy for socio-economic class which seems to often be more a predictor of poorer outcomes. The CDC data is also quite dated being from 2010
- Measured performance with HEDIS data
- Developer cited evidence which demonstrates racial disparity in testing, age disparitiy, gender disparity
- HEDIS data from 2014 to 2016 is included. Comprehensive diabetes care continues to be a HEDIS measure in 2018. Studies are cited in the original application that compare achieving the A1C testing goal between age groups, sex, education levels, and race/ethnicity.
- There is reported disparities with Hispanic patients with diabetes having lower rates of a 1C testing (63% versus 73%)
- There is still a need for improvement. Medicare reports the highest rate of performance at 93%, commercial plans report the lowest at 89%
- These plans didn't report ethnicity data, but cited CDC who reported Hispanic patients having the lowest rate of performance.
- Women have higher performance rates than men.

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.

Composite measures only:

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

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

Evaluation of Reliability and Validity (and composite construction, if applicable): **Staff evaluation**

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] or [Scientific Methods Panel] is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

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

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

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

Committee pre-evaluation comments

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

Reliability

- Data elements are clearly defined. No concerns with the code or value set. The measure should be able to be consistently implemented.
- No concerns about reliability specifications.
- Concur with the analysis of the staff evaluator.
- No concerns
- No need to discuss and vote
- Reliability of HEDIS measures are discussed in the preceding application.
- No concerns about reliability and I agree with the staff that we do not need to vote on reliability

Reliability Testing

- I have no concerns related to the reliability of this measure. With that in mind, there is no need to discuss and vote again on reliability.
- No concerns with the reliability of the measure.
- No concerns about validity.
- Concur with the analysis of the staff evaluator.
- No concerns
- No concerns about reliability and I agree with the staff that we do not need to vote on reliability

Validity Testing

- I have no concerns on the validity of this measure. With that in mind, there is no need to discuss and vote again on validity.
- The Developer provided results over three years of HEDIS results for commercial, Medicare and Medicaid populations. Results are used to compare health plans, and can be used to compare providers. Results are used by States and by CMS in incentive programs. No concerns with missing data. The measure is collected by chart review and claims.
- No concerns.
- Concur with the analysis of the staff evaluator
- No concerns; they provide construct and face validity testing;
- Validity of HEDIS measures are discussed in the preceding application.
- No concerns
- No concerns about validity and I agree with the staff that we do not need to vote on validity

Other threats

- The exclusions are consistent with the evidence. Patient groups are appropriate. Hospice and individuals without a diagnosis of diabetes in measurement year or year prior are excluded.
- No threats to Validity from Exclusions or Risk Adjustment
- Concur with the analysis of the staff evaluator.
- NA
- N/a
- No concerns about validity and I agree with the staff that we do not need to vote on validity

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 measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records). While only some data elements are in defined fields in electronic sources, the elements are generated as byproduct of care processes. This measure is also a HEDIS measure and NCQA conducts audits to verify that HEDIS specifications are met.
- This is not an eMeasure.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- This is a process measure for a test that is reimbursable by payers. As such, what is being measured is a byproduct of the care delivery and reimbursement process.
- Measure uses claims and chart review to collect the data necessary for reporting. It is resource intensive and more expensive to report than administrative measures. As electronic health records become more accessible for measure reporting it will become less burdensome to report.
- Data elements are routinely generated during usual care delivery. The data elements are available in an electronic form This is not an eMeasure. Preliminary rating Moderate
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one.
- No concerns; data already being collected in feasible manner
- The HEDIS Audit process is described in the initial applications.
- Easily feasible to collect this data either from EHR or from claims data.
- The measure is extremely feasible. A1c is routine/easily measured, easily retrieved form the EHR and billing/administrative data.

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

- HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
- STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
- INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States.
- ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.
- DIABETES RECOGNITION PROGRAM: This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care.
- QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial, Medicaid, and Medicare health plans.

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

- NCQA publishes HEDIS results annually in its Quality Compass tool. The measure receives feedback through the Policy Clarification Support System. This is a long-standing, well-understood measure so NCQA receives very few questions or requests for clarification about it. Questions received through the Policy Clarification Support system have generally centered around clarification on what types of HbA1c laboratory tests qualify for numerator compliance. Feedback has not required modification to the measure.

Additional Feedback: NA

***Questions for the Committee**:*

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

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

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

4b. Usability 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

- Since 2010, this measure has reported stable and high levels of performance on average (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states reported data on this measure. These data are nationally representative.

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

Unexpected findings (positive or negative) during implementation

- The developer did not report any unexpected findings.

Potential harms

- The developer did not report any unintended consequences.

Additional Feedback:

***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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- This measure is widely used in public reporting and accountability programs.
- HEDIS is the primary tool used for measurement. Health plans and providers are able to provide feedback on measure results during the reporting process.
- The measure is being reported publicly. Seven accountability programs are cited by the developer. NCQA publishes HEDIS results annually in its Quality Compass tool. The measure receives feedback through the Policy Clarification Support System. Questions have generally centered around clarification on what types of HbA1c laboratory tests qualify for numerator compliance. Feedback has not required modification to the measure.
- How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Given the critical role that A1C measurement seems to play in diabetes, one would expect a robust engagement with such a large identified community of patients. Ultimately patients are the "measured" entity.
- No concerns; publically reported and already part of HEDIS
- Also reported in NCQA report cards, NCQA State of Health Care Annual report, California PFP, NCQA ACO and Health Plan accreditation, NCQA DRP and Quality Compass
- The original application listed the numerous publications and types of providers that reference HEDIS data.
- Publicly reported. Health plans and providers able to obtain performance data.

Usability

- Since this measure assesses whether A1c testing has been done and does not assess attainment of a specific threshold for glycemic control, there should be no unintended consequences of this particular measure. If anything, this measure can serve as the basis of individualized, patient centric care where patients and clinicians use this information to assess attainment of individualized goals.
- No identified unintended consequences. Has an impact on quality of life.
- There is need for improvement. The harms can be minimized by avoiding hypoglycemia, A1c levels below 8 in older patients with co-morbidities. The A1c does not reflect episodes of hypoglycemia, nor does it give an indication of the time glucose is in a normal, near normal or acceptable range. .
- "There are many great examples of how these outcomes are communicated to providers but fewer on how these data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- Many diabetic patients speak of the challenges with A1C Targets – I find no issue with the measure, this value is important. However, how this value is used is a byproduct of the measure and should be considered – many patients struggle to meet these targets and sacrifice quality of life to do so. Language or engagement with patients on this issue may help to provide more guidance on how to word this guidance on measure implementation.

Criterion 5: Related and Competing Measures

Related or competing measures

- NA

Harmonization

- NA

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: 0057
Measure Title: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Testing

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

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

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

   ☒Yes (go to Question #2)

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

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

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

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

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☒High (go to Question #6)

   ☐Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.

   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

   REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   ***TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
   ☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:
    ☐High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)
    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)
    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete]
    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

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

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

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

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*

REFERENCE: *Testing attachment, section 2b1.*

TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

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

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☒Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0057

Measure Title: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Testing

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

Date of Submission: 4/9/2018

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: Receiving a HbA1c test during the measurement year

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Adults with diabetes (type 1 or 2) >>> HbA1c test is performed>>> Test results are evaluated>>>HbA1c Health provider determines treatment to keep HbA1c at desirable level>>>Maintenance or improvement in HbA1c level and/or quality of life (desired outcome).

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.)

N/A

****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.

N/A

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

Table 1. American Diabetes Association Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Diabetes Care (American Diabetes Association) Standards of Medical Care in Diabetes-2018. Diabetes Care January 2018. 41 (Supp 1): S55-64. https://doi.org/10.2337/dc18-S006 URL: http://diabetesed.net/wp-content/uploads/2017/12/2018-ADA-Standards-of-Care.pdf 2013 Diabetes Care (American Diabetes Association) Standards of Medical Care in Diabetes-2013. Diabetes Care January 2013 36:S1-e4; doi: 10.2337/dc13-S001 URL: http://mcintranet.musc.edu/agingq3/calculationswesbite/ADA%20Guidelines/ADA%20Binder.pdf

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	Recommendations (2018) • Perform the A1C test at least two times a year in patients who are meeting treatment goals (and who have stable glycemic control). E • Perform the A1C test quarterly in patients whose therapy has changed or who are not meeting glycemic goals. E • Point-of-care testing for A1C provides the opportunity for more timely treatment changes. E Recommendations (2013) Same as above
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Level of Evidence & Description: E Expert consensus or clinical experience 2013 Same as above
Provide all other grades and definitions from the evidence grading system	2018 Level of Evidence & Description: A Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., the "all or none" rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis B Supportive evidence from well-conducted cohort studies, including: • Evidence from a well-conducted prospective cohort study or registry • Evidence from a well-conducted meta-analysis of cohort studies • Supportive evidence from a well-conducted case-control study C Supportive evidence from poorly controlled or uncontrolled studies, including:

	• Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) • Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation 2013 Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. 2013 Same as above
Provide all other grades and definitions from the recommendation grading system	2018 No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. 2013 Same as above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The ADA does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on two other systematic reviews that support the ADA's recommendations in Table 3.
Estimates of benefit and consistency across studies	See Table 3.

What harms were identified?	See Table 3.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	There have been no new studies that contradict the current body of evidence.

Table 2. American Geriatric Society Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 American Geriatrics Society (AGS). Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. American Geriatrics Society Panel on the Care for Older Adults with Diabetes Mellitus. Journal of American Geriatric Society. 2013 November; 61 (11): 2020-2026. Doi:10.1111/jgs.12514 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064258/pdf/nihms583558.pdf 2013 American Geriatric Society (AGS) Guidelines for Improving Care of the Older Adults with Diabetes Mellitus. California Healthcare Foundation/American Geriatric Society Panel on Improving Care for Elders with Diabetes. American Geriatric Society. May 2003 – 51 (5) Supplement, JAGS. URL: http://www.medicine.emory.edu/ger/bibliographies/geriatrics/bibliography87_files/Guidelines_for_ Improving_the_Care_of_the_Older_Person_with_Diabetes_Mellitus.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the	Recommendations (2018) Pg. 4 “Glycemic Control” General Recommendations 1. Target goal for glycosylated hemoglobin (HbA1c) in older adults generally should be 7.5% to 8%. HbA1c between 7% and 7.5% may be appropriate if it can be safely achieved in healthy older adults with few comorbidities and good functional status. Higher HbA1c targets (8–9%) are appropriate for older adults with multiple comorbidities, poor health, and limited life expectancy. (1A

conclusions from the SR.	evidence for HbA1c 7–8%, and IIA for 8–9%) There is potential harm in lowering HbA1c to less than 6.5% in older adults with type 2 DM. (IIA) 2. Older adults with DM whose individual targets are not being met should have their HbA1c levels measured at least every 6 months and more frequently as needed or indicated. For older adults with stable HbA1c over several years, measurement every 12 months may be appropriate. (IIIB) Recommendations (2013) Pg. S270 "Glycemic Control" General Recommendations 1. For older persons, target hemoglobin A1c (A1C) should be individualized. A reasonable goal for A1C in relatively healthy adults with good functional status is 7% or lower. For frail older adults, persons with life expectancy of less than 5 years, and others in whom the risks of intensive glycemic control appear to outweigh the benefits, a less stringent target such as 8% is appropriate. (IIIB)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Quality of Evidence • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytic studies, or from multiple time-series studies, or from dramatic results in uncontrolled experiments • Level III: Evidence from respected authorities, based on clinical experience, descriptive studies, or reports of expert committee Strength of Evidence • A: Good evidence to support the use of a recommendation; clinicians should do this all the time • B: Moderate evidence to support the use of a recommendation; clinicians should do this most of the time 2013 Same as above
Provide all other grades and definitions from the evidence grading system	2018 Quality of Evidence • Level I: Evidence from at least one properly designed randomized, controlled trial • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytic studies, or from multiple time-series studies, or from dramatic results in uncontrolled experiments Strength of Evidence

	• A: Good evidence to support the use of a recommendation; clinicians should do this all the time • B: Moderate evidence to support the use of a recommendation; clinicians should do this most of the time • C: Poor evidence to support or to reject the use of a recommendation; clinicians may or may not follow the recommendation • D: Moderate evidence against the use of a recommendation; clinicians should not do this • E: Good evidence against the use of a recommendation; clinicians should not do this 2013 Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. 2013 Same as above
Provide all other grades and definitions from the recommendation grading system	2018 No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. 2013 Same as above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AGS does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the AGS systematic review, we provide information on two other systematic reviews that support the AGS's recommendations in Table 3.
Estimates of benefit and	See Table 3.

consistency across studies	
What harms were identified?	See Table 3.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	There have been no new studies that contradict the current body of evidence.

Table 3. Systematic Review

Citation	Department of Veteran Affairs, Department of Defense. VA/DoD clinical practice guideline for the management of diabetes mellitus. 2010. Washington (DC): Department of Veteran Affairs, Department of Defense. Retrieved from http://www.healthquality.va.gov/diabetes/DM2010_FUL-v4e.pdf.
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	The evidence for this measure is based on measurement of blood glucose or HbA1c to facilitate glycemic control in adults with diabetes. Monitoring of blood glucose can be conducted by patients through self-monitoring (SBGM) or by the provider through point of care treatment (PoCT). Self-monitoring includes using at home blood glucose tests to continuously measure glucose levels. HbA1c tests are conducted or ordered by a provider to measure the average blood glucose over a three-month period. Results from monitoring assist providers and patients with maintaining or improving glycemic control and reducing complications from diabetes.
Grade assigned for the quality of the quoted evidence with definition of the grade	Level of Evidence (LE) I At least one properly done RCT II-1 Well designed controlled trial without randomization II-2 Well designed cohort or case-control analytic study II-3 Multiple time series, dramatic results of uncontrolled experiment Quality of Evidence Fair High grade evidence (I or II-1) linked to intermediate outcome; or Moderate grade evidence (II-2 or II-3) directly linked to health outcome Strength of Recommendation B A recommendation that the intervention may be useful/effective C A recommendation that the intervention may be considered

Provide all other grades and associated definitions of the evidence in the grading system	Level of Evidence (QE) III Opinion of respected authorities, case reports, and expert committees Quality of Evidence Good High grade evidence (I or II-1) directly linked to health outcome Poor Level III evidence or no linkage of evidence to health outcome Strength of Recommendation A A strong recommendation that the intervention is always indicated and acceptable D A recommendation that a procedure may be considered not useful/effective, or may be harmful. I Insufficient evidence to recommend for or against – the clinician will use clinical judgment Net Effect of the Intervention Substantial: More than a small relative impact on a frequent condition with a substantial burden of suffering; *or* A large impact on an infrequent condition with a significant impact on the individual patient level. Moderate: A small relative impact on a frequent condition with a substantial burden of suffering; *or* A moderate impact on an infrequent condition with a significant impact on the individual patient level. Small: A negligible relative impact on a frequent condition with a substantial burden of suffering; *or* A small impact on an infrequent condition with a significant impact on the individual patient level. Zero or Negative: Negative impact on patients; or No relative impact on either a frequent condition with a substantial burden of suffering; *or* An infrequent condition with a significant impact on the individual patient level.
What is the time period covered by the body of evidence?	1997-2008
Quantity and Quality of	Periodic HbA1c measurements: over 20 studies including 14 RCTs, 4 descriptive prospective studies, 1 comparative retrospective study, clinical trials, observational studies, epidemiological data, and literature reviews.

Body of Evidence	Instruction in interpretation and use of SBGM: over 20 RCTs, clinical trials, and literature reviews SBGM in non-insulin requiring type 2 diabetics to adjust treatment: over 20 studies including RCTs Utilizing remote SBGM data: over 40 RCTs
What is the overall quality of evidence <u>across studies</u> in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are over 100 studies in the evidence review that examine the effectiveness of measuring HbA1c or blood glucose and glycemic control. The evidence for periodic HbA1c measurements is strong. The VA/DoD evidence review gave this recommendation the following grading: LE=II, QE=fair, SR=B. The fair rating for the quality of evidence (see quality grading) indicates that the evidence can be linked to the health outcome. The B grading for this evidence signifies that HbA1c testing may be useful or effective. Furthermore, the level of evidence indicates that the studies used were well designed controlled trials, cohort or case controlled studies, or included multiple time series.
Estimates of benefit and consistency across studies in body of evidence – what are the estimates of benefits?	Randomized clinical trials have demonstrated that improved glycemic control, as evidenced by reduced levels of glycohemoglobin, correlates with a reduction in the development of microvascular complications in both Type 1 and Type 2 diabetes (DCCT 1993, Ohkubo 1995). In particular, the Diabetes Control and Complications Trial (DCCT) showed that for patients with Type 1 diabetes mellitus, important clinical outcomes such as retinopathy (an important precursor to blindness), nephropathy (which precedes renal failure), and neuropathy (a significant cause of foot ulcers and amputation in patients with diabetes) are directly related to level of glycemic control (DCCT 1993). Similar reductions in complications were noted in a smaller study of intensive therapy of patients with Type 2 diabetes by Ohkubo and co-workers, which was conducted in the Japanese population (Ohkubo 1995). Based primarily on the strength of the DCCT study and the corroborating evidence, most experts agree that control of glycemia as measured by glycohemoglobin is an important way to minimize the incidence of the microvascular complications of diabetes (ADA 2013). Consequently, based on the findings of the DCCT and UKPDS, many organizations in this country published guidelines for the achievement of good metabolic control in diabetes (ADA 2013). American Diabetes Association. Standards of Medical Care in Diabetes—2013 Diabetes Care January 2013 36:S11-S66; doi:10.2337/dc13-S011 The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes and progression of long-term complications in insulin-dependent mellitus. N Engl J Med 329:977-86, 1993.
What harms were studied and how do	No harms associated with testing were identified in the evidence reviewed. However, there are potential harms that may stem from a program of Hba1c testing followed by tight control. This tight glycemic control may result in episodes of

they affect the net benefit (benefits over harms)?	hypoglycemia. One study concludes that intensive glycemic control does not seem to reduce all-cause mortality in patients with type 2 diabetes. Intensive glycemic control increases the relative risk of severe hypoglycemia by 30% (Hemmingsen et al. 2011). Hemmingsen, B. et al. Intensive glycemic control for patients with type 2 diabetes: systematic review with meta-analysis and trial sequential analysis of randomized clinical trials. BMJ 2011; 343:d6898. https:// http://dx.doi.org/10.1136/bmj.d6898
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous studies have been conducted since the systematic reviews we cite in this table, none of which change the conclusion that routine HbA1c testing for individuals with diabetes are appropriate. Below we list two additional studies that support this measure. Perrotta PL, Jones R, Souers RJ, et al. Frequency of monitoring hemoglobin A1c, low density lipoprotein and urine protein laboratory testing. Archives of Pathology and Laboratory Medicine 2014; 138:1009-1014. doi: 10.5858/arpa.2013-0349-CP Driskell OJ, Holland D, Waldron JL, et al. Reducing testing frequency for glycated hemoglobin HbA1c, is associated with deteriorating diabetes control. Diabetes Care 2014; 37(10):2731-2737. https://doi.org/10.2337/dc14-0297

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.

N/A

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

N/A

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

N/A

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0057
Corresponding Measures:
De.2. **Measure Title:** Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Testing
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received an HbA1c test during the measurement year.
1b.1. **Developer Rationale:** Testing hemoglobin A1c levels in patients with diabetes is an important component of diabetes treatment and care. Results from these tests aids clinicians in providing patients with optimal treatment that will maximize diabetes control and in turn prevent complications of diabetes that threaten to impact quality of life.

S.4. **Numerator Statement:** Patients who had an HbA1c test performed during the measurement year.
S.6. **Denominator Statement:** Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.
S.8. **Denominator Exclusions:** Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Members who do not have a diagnosis of diabetes in any setting, during the measurement year or the year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes in any setting, during the measurement year or the year prior to the measurement year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data, Paper Medical Records
S.20. **Level of Analysis:** Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

IF this measure is included in a composite, NQF Composite#/title:
0731:Comprehensive Diabetes Care

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_0057_HbA1c_Testing_7.1.docx
1a.1 <u>For Maintenance of Endorsement:</u> 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.
No

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.
Testing hemoglobin A1c levels in patients with diabetes is an important component of diabetes treatment and care. Results from these tests aids clinicians in providing patients with optimal treatment that will maximize diabetes control and in turn prevent complications of diabetes that threaten to impact quality of life.

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.*
HEDIS
The following data are extracted from HEDIS data collection reflecting the most recent years of reporting for this measure. Performance data is summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year and product line (i.e. commercial, Medicaid, and Medicare).

Comprehensive Diabetes Care: HbA1c Testing
N= Number of plans reporting

Commercial Rate
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2014| 412| 89.42%| 3.73%| 73.61%| 85.19%| 87.13%| 89.55%| 92.07%| 97.83%
2015| 126| 89.47%|5.79%| 0.00%| 85.40%| 87.76%| 90.08%| 92.15%| 94.16%| 97.33%
2016| 413| 89.91%| 3.53%| 73.77%| 85.49%| 87.87%| 90.24%| 92.21%| 94.34%| 97.08%

Medicaid Rate
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2014| 223| 86.31%| 4.84%| 70.80%| 80.29%| 83.19%| 86.20%| 89.55%| 91.94%| 98.63%
2015| 263| 85.95%| 5.34%| 65.22%| 79.56%| 82.98%| 85.95%| 89.43%| 92.88%| 100.00%
2016| 270| 86.66%| 5.66%| 49.37%| 80.95%| 84.32%| 87.10%| 90.05%| 92.78%| 100.00%

Medicare Rate
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2014| 475| 92.72%| 4.17%| 60.00%| 89.05%| 91.06%| 93.27%| 95.30%| 96.80%| 98.97%
2015| 461| 93.10%| 3.81%| 61.43%| 89.29%| 91.84%| 93.69%| 95.22%| 96.84%| 100.00%
2016| 472| 93.54%| 3.38%| 61.69%| 89.61%| 91.97%| 93.99%| 95.62%| 97.07%| 100.00%

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.**
N/A

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.

HEDIS data is stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escare et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escare 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership were designed to promote standardized methods for collecting these data. These measures follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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

Between 2000 and 2010, the CDC monitored the percentage rates of diabetic adults over age 18 that received two or more hemoglobin A1c (HbA1c) tests within the last year. The rates were monitored based on race/ethnicity, sex, age, and education level. The most recent data shows Hispanics with the lowest rates of HbA1c tests at 63.7% (CDC, 2012). Whites and Blacks had nearly equal percentages at 73.4% and 73.1%, respectively (CDC, 2012). The most recent data for the percentage of A1c testing, based on education level, was highest in people with an education above high school (72.8%) (CDC, 2012). The percentage of high school-educated adults with A1c tests was 67.5% and 54.4% in adults with less than a high school education (CDC, 2012). Since 2000, diabetic adults who attained higher than a high school degree have consistently reported having frequent number of A1c tests, compared to diabetic adults with less than a high school education. However, in 2008 and 2009, more diabetic adults with less than a high school education reported having frequent A1c testing compared to high school educated adults (CDC, 2012). In 2010, the highest percentage of A1c tests based on age was seen in the 65-74 age group (78.2%), followed by the over 75 age group (75.7%) (CDC, 2012). Adults in the 45-64 and 18-44 age groups had the lowest percentages for A1c tests at 72.4% and 63.5%, respectively (CDC, 2012). Based on data from the CDC, women receive A1c tests at higher rates than men. In 2010, 74% of diabetic women had two or more A1c tests in the past year (CDC, 2012). Only 71.5% of men had two or more A1c tests in 2010.

Centers for Disease Control and Prevention. 2012. CDC's Diabetes Program-Data and Trends-Prevalence of Diabetes-Percent of 2 or More A1c Tests in the Last Year for Adults Aged = 18 Years by Race/Ethnicity. Retrieved from https://www.cdc.gov/diabetes/statistics/preventive/fy_ac1test.htm.

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

Measure Number (*if previously endorsed*)**:** 0057
Measure Title: Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Testing
Date of Submission: 3/1/2018
Type of Measure:

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

Instructions

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

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

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

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

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]
AND
If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

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

☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Electronic health data, electronic clinical data: laboratory	☐ other: Click here to describe

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

1.3. What are the dates of the data used in testing? 2010 - 2012

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

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

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
Health Plan Level
We calculated the measure score reliability and construct validity from HEDIS data that included 418 commercial health plans, 500 Medicare health plans, and 201 Medicaid health plans. The sample included all commercial, Medicare, and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

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*)
2012 data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible plans for the measure across health plans.

HEDIS Health Plan

Product Type	Number of Plans	Median Number of Eligible Patients per Plan
Commercial HMO	219	2,599
Commercial PPO	199	6,476
Medicaid HMO	201	1,774
Medicare HMO	349	1,586
Medicare PPO	151	1,527

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
Reliability:
Reliability of the health plan measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Validity:
Validity of the health plan measure was demonstrated through construct validity using the entire HEDIS data sample (described above) and through a systematic assessment of face validity with expert panels.

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

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*)
Reliability Testing of Performance Measure Score:
Reliability was estimated by using the beta-binomial model for the health plan measure. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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*)
Health Plan Level - 2012

Product Type	Reliability per Beta Binomial Model
Commercial	0.98
Medicare	0.96
Medicaid	0.97

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?*)
Health Plan Level
The values for the beta-binomial statistic across all product lines for the health plan level measure suggest the measure has high reliability.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Method of Testing Construct Validity – Health Plan Level

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality focused on diabetes care hypothesized to be related, which are listed below.

- Hemoglobin (HbA1c) Poor Control (>9%)
- HbA1c Good Control (<8%)
- Eye Examinations (Eye Exams)
- Medical attention for nephropathy

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of Assessing Face Validity – Health Plan Level

We describe below NCQA's process for both measure development and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assure that the measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is

vetted by NCQA's MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.
STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed periodically, based on changes in evidence and guidelines. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support (PCS) portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures. Over the past four years, NCQA has received and responded to an average of 39 inquiries per year on this measure.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2.

See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
The results from construct validity testing of the health plan level measure are presented by product line in Tables 1a, 1b, and 1c below.

Table 1a. Correlations among Diabetes Measures in Commercial Health Plans - 2012

	Pearson Correlation Coefficients			
	CDC – Medical Attention for Diabetic Nephropathy	HbA1c Poor Control (>9.0%)	HbA1c Control (<8.0%)	Eye Exams
HbA1c Testing	0.76	-0.67	0.66	0.69

Note: All correlations are significant at p<0.0001

Table 1b. Correlations among Diabetes Measures in Medicare Health Plans - 2012

	Pearson Correlation Coefficients			
	CDC – Medical Attention for Diabetic Nephropathy	HbA1c Poor Control (>9.0%)	HbA1c Control (<8.0%)	Eye Exams
HbA1c Testing	0.43	-0.62	0.62	0.60

Note: All correlations are significant at p<0.0001

Table 1c. Correlations among Diabetes Measures in Medicaid Health Plans - 2012

	Pearson Correlation Coefficients			
	CDC – Medical Attention for Diabetic Nephropathy	HbA1c Poor Control (>9.0%)	HbA1c Control (<8.0%)	Eye Exams
HbA1c Testing	0.56	-0.64	0.61	0.53

Note: All correlations are significant at p<0.0001

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?*)

Construct Validity

Across all product lines, the correlations are moderate to strong and statistically significant. These results confirmed the hypothesis that the diabetes measures are correlated with each other. Coefficients with absolute value of less than .3 are generally considered indicative of weak associations. Absolute values of .3 to .59 are considered moderate associations, absolute values of .6 to .69 indicate a strong positive relationship, and absolute values of .7 or higher indicate a very strong positive relationship. These correlation results suggest that at the plan level the measure has sufficient validity.

Note: Correlation values with the HbA1c Poor Control measure are all negative because it is a "lower is better quality" measure, while the other measures are all "higher is better".

Face Validity

NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that the *CDC - HbA1c Testing* measure is measuring what it intends to measure. The results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)

Testing was not performed for the excluded sample.

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*)
Testing was not performed for the excluded sample.

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*)
Testing was not performed for the excluded sample.

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

2b3.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

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

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**
☐ **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*)
N/A

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

_____

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample, size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans performance is significantly different from each other.

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*)

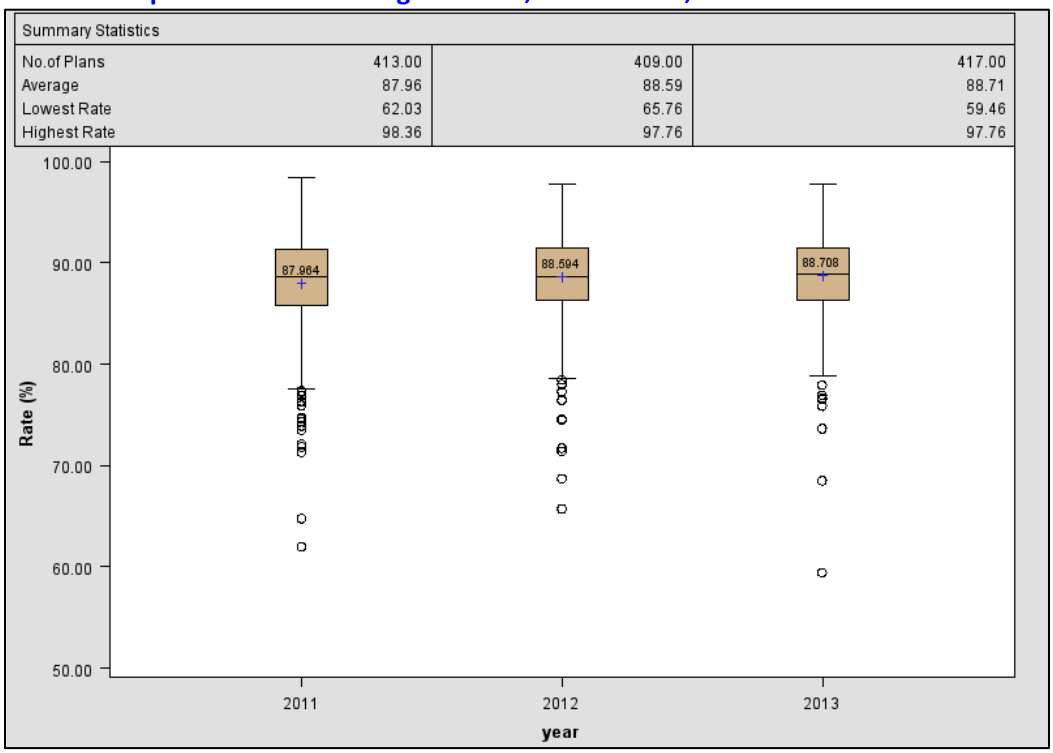
Health Plan - 2012

Product Type	N	Mean (%)	St Dev (%)	P10th (%)	P25th (%)	P50th (%)	P75th (%)	P90th (%)	IQR (%)	P value
Commercial HMO	219	90.09	4.05	85.64	87.59	90.54	92.88	94.92	5.00	<0.05

Commercial PPO	199	87.17	4.18	82.09	84.84	87.74	89.98	91.48	5.00	<0.05
Medicaid HMO	201	82.99	6.12	75.91	79.23	83.21	87.32	90.97	8.00	<0.05
Medicare HMO	349	91.45	4.15	86.62	89.29	91.90	94.16	96.06	5.00	<0.05
Medicare PPO	151	91.00	3.18	87.62	89.29	91.24	92.59	94.54	3.00	<0.05

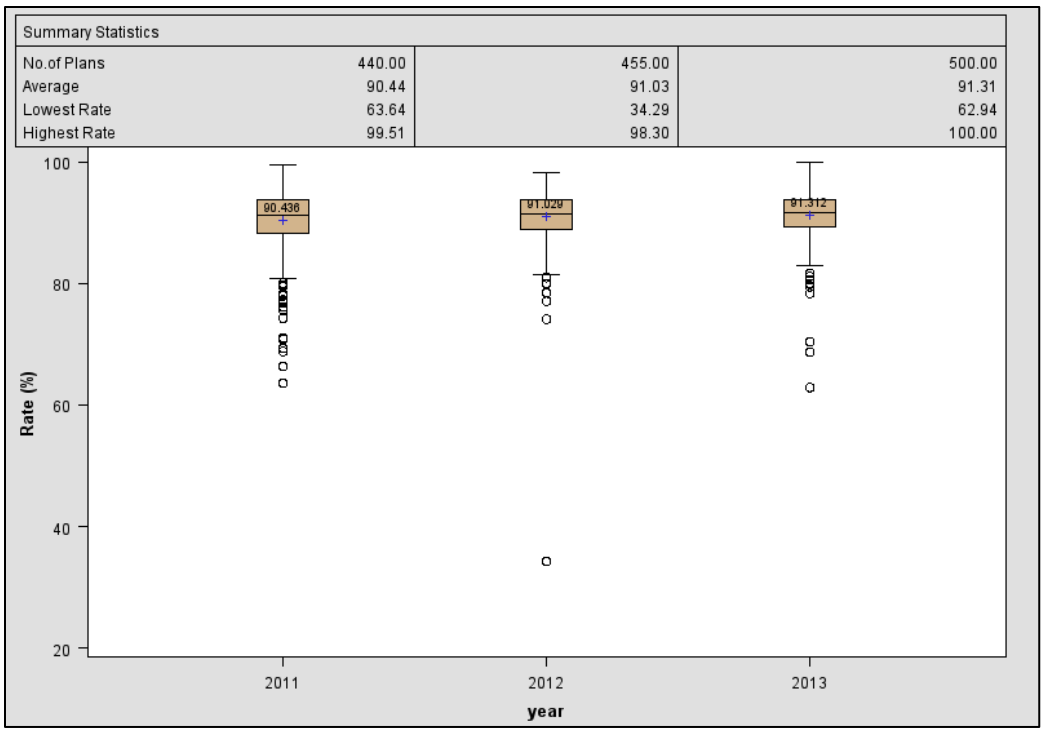
N = total number of plans reporting data
IQR: Interquartile range
p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

Chart 1. Boxplot of HbA1c Testing Measure, Commercial, HEDIS 2011-2013*



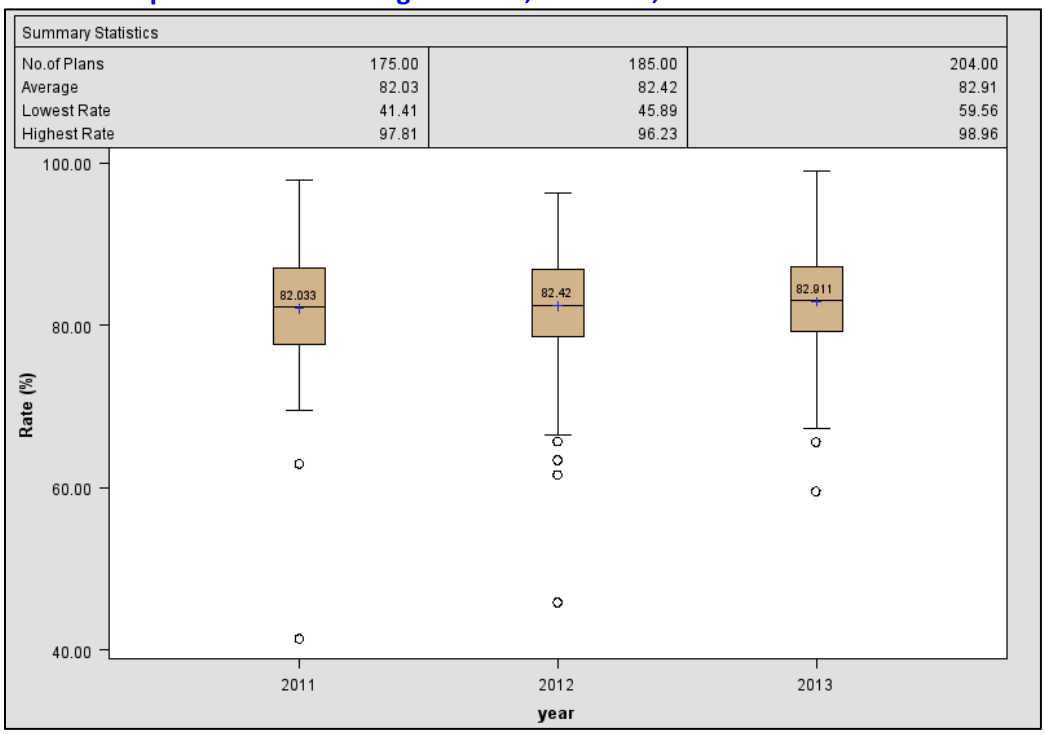
* In this chart, data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Chart 2. Boxplot of HbA1c Testing Measure, Medicare, HEDIS 2011-2013*



* In this chart, data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012.

Chart 3. Boxplot of HbA1c Testing Measure, Medicaid, HEDIS 2011-2013*



* In this chart, data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012.

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?*)

Across all product lines, the difference between the 25th (better performance) and 75th percentile is statistically significant. Overall, these results suggest there are meaningful differences in performance.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

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*)
This measure is collected with a complete sample.

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

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):*
Endocrine, Endocrine : Diabetes

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):*
Populations at Risk

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.)*
N/A

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0057_CDC_HbA1c_Testing_Value_Set_.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
No

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

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients who had an HbA1c test performed during the measurement year.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
ADMINISTRATIVE CLAIMS: An HbA1c test (HbA1c Tests Value Set) performed during the measurement year, as identified by claim/encounter or automated laboratory data.

Due to the extensive volume of codes associated with identifying numerator events for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD: At a minimum, documentation in the medical record must include a note indicating the date when the HbA1c test was performed and the result or finding. Count notation of the following in the medical record:

- A1c.
- HbA1c
- HgbA1c.
- Hemoglobin A1c.
- Glycohemoglobin A1c.
- Glycohemoglobin.
- Glycated hemoglobin.
- Glycosylated hemoglobin.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients with diabetes can be identified two ways:

-CLAIM/ENCOUNTER DATA: Patients who had two face-to-face encounters, in an outpatient setting or nonacute inpatient setting, or ED setting on different dates of service, with a diagnosis of diabetes, or one face-to-face encounter in an acute inpatient, with a diagnosis of diabetes, during the measurement year or the year prior to the measurement year. Organizations may count services that occur over both years.
*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

-PHARMACY DATA: Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year.

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES (TABLE CDC-A):
Alpha-glucosidase inhibitors:
Acarbose, Miglitol
Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metaformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metaformin-saxagliptin, Metformin-sitagliptin, Sitagliptin-simvastatin

Insulin:
Insulin aspart, Insulin aspart-insulin aspart protamine, insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitagliptin

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Members who do not have a diagnosis of diabetes in any setting, during the measurement year or the year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes in any setting, during the measurement year or the year prior to the measurement year.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
ADMINISTRATIVE CLAIMS:
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD:
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and had a diagnosis of polycystic ovaries any time in the patient's history through December 31 of the measurement year.
OR
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and*

coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)
N/A

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 = Higher 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.)*
STEP 1. Determine the eligible population. To do so, identify patients who meet all the specified criteria.
-AGES: 18-75 years as of December 31 of the measurement year.
-EVENT/DIAGNOSIS: Identify patients with diabetes in two ways: by claim/encounter data and by pharmacy data.

Claim/Encounter Data:
-Patients who had at least two outpatient visits, observation visits or nonacute inpatient encounters on different dates of service, with a diagnosis of diabetes. Visit type need not be the same for the two visits.
-Patients with at least one acute inpatient encounter with a diagnosis of diabetes.
-Patients with at least one ED visit with a diagnosis of diabetes.
*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

Pharmacy Data:
Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year. *SEE PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES IN S.7

STEP 2. Determine the number of patients in the eligible population who had a recent HbA1c test during the measurement year through the search of administrative data systems.
STEP 3. Identify patients with a most recent HbA1c test performed.
STEP 4. Identify the most recent HbA1c test with result (numerator compliant). Identify a missing result or no HbA1c test done during the measurement year (not numerator compliant).
STEP 5. Exclude from the eligible population patients from step 2 for whom administrative system data identified an exclusion to the service/procedure being measured. *SEE DENOMINATOR EXCLUSION CRITERIA IN QUESTION S.8
STEP 6. Calculate the rate (number of patients that had an HbA1c test).

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Electronic Health Records, Paper Medical 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.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

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.)*
N/A

2. Validity – See attached Measure Testing Submission Form
NQF_Testing_0057_HbA1c_Testing_7.1.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.
No

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.
No

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.
No - This measure is not risk-adjusted

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
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**.
Some data elements are in defined fields in electronic sources

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).
To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

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

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds

immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
N/A

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 Health Plan Rating http://reportcard.ncqa.org/plan/external/plansearch.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Health Plan Rating http://reportcard.ncqa.org/plan/external/plansearch.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Payment Program IHA California Pay for Performance http://www.iha.org/manuals_operations_2014.html Regulatory and Accreditation Programs NCQA Accreditation http://www.ncqa.org/tabid/123/Default.aspx Accountable Care Organizations (ACO) http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx NCQA Accreditation http://www.ncqa.org/tabid/123/Default.aspx Accountable Care Organizations (ACO) http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Professional Certification or Recognition Program NCQA Diabetes Recognition Program

	http://www.ncqa.org/Programs/Recognition/DiabetesRecognitionProgramDRP.aspx Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality: http://www.ncqa.org/tabid/836/Default.aspx

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

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

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population.

HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rankings which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

DIABETES RECOGNITION PROGRAM: This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care. The DRP Program has 11 measures which cover other areas such as: HbA1c Control, Blood Pressure Control, LDL Control, Eye Examinations, Nephropathy Assessment, Smoking and Tobacco Use and Cessation advice or treatment. Eligible clinicians will abstract data from the charts of diabetes patients (25 patients for a single applicant) and submit this information to NCQA for review.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial, Medicaid, and Medicare health plans. As of Fall 2017, a total of 184 Medicare Advantage health plans were accredited using this measure among others covering 9.2 million Medicare beneficiaries; 451 commercial health plans covering 113 million lives; and 125 Medicaid health plans covering 35 million lives. Health plans are scored based on performance compared to benchmarks.

INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of eight health plans representing 10 million insured persons. IHA is responsible for collecting data, deploying a common measure set, and reporting results for approximately 35,000 physicians in nearly 200 physician groups. This program represents the longest running U.S. example of data aggregation and standardized results reporting across diverse regions and multiple health plans. California consumers benefit from the availability of standardized performance results from a common measure set, which are available to the public through the State of California, Office of the Patient Advocate

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?*)

N/A

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.*)
N/A

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
This is a long-standing, well-understood measure so NCQA receives very few questions or requests for clarification about it. Questions received through the Policy Clarification Support system have generally centered around clarification on what types of HbA1c laboratory tests qualify for numerator compliance.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the Annual State of Healthcare Quality and the Qualified Health Plan Quality Rating System.

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

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.

Since 2010, this measure has reported stable and high levels of performance on average (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states reported data on this measure. These data are nationally representative.

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.**

There were no identified unexpected findings during testing or since implementation of this measure.

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

There were no identified unexpected benefits during testing or since implementation of this measure.

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.

No

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?

No

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

N/A

5b. Competing Measures

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

OR

Multiple measures are justified.

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

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

N/A

Appendix

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

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.

DIABETES EXPERT PANEL:
Bill Herman (Chair), MD, Univ. of Michigan Health System
David Aron, MD, Department of Veteran's Affairs
James Fain, PhD, RN, University of Massachussetts
Jerry Cavallerano, OD, Beetham Eye Institute
John Thompson, MD, Retina Specialists
Judith Fradkin, MD, NIDDK/NIH
Lynne Levitsky, MD, Massachusetts General Hospital
Mark Cziraky, PharmD, Healthcore
Richard Hellman, MD, Private Practice, Diabetes & Endocrinology
Seth Rubenstein, DPM, Reston Hospital Center, INOVA Fair Oaks Hospital
Stephen Fadem, MD, Baylor College of Medicine
Ted Ganiats, MD, Univ. of California, San Diego
Nancy Van Vessem, MD, Capital Health Plan

HEDIS EXPERT CODING PANEL
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Aqurate Health Data Management, Inc.
Nelly Leon-Chisen, RHIA, American Hospital Association
Alec McLure, MPH, RHIA, CCS-P, Verscend Technologies
Michele Mouradian, RN, BSN, Change HealthCare
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, WellCare Health Plans, Inc.

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management

Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

CLINICAL PROGRAMS COMMITTEE
Randall Curnow, MD, MBA, FACP, FACHE, FACPE (Chair), TriHealth
Suzanne Berman, MD, FAAP, Plateau Pediatrics
Brooks Daveman, MPP, Tennessee Division of Health Care Finance and Administration
Marcus Friedrich, MD, MBA, FACP, New York State Department Health Empire State Plaza, Coming Towne
Jennifer Gutzmore, MD, Cigna
Melissa Hogan, MPH, Aon
Adriana Matiz, MD, FAAP, Ambulatory Care Network
Lisa Morrise, Marts, LAM Professional Services, LLC
Deborah Murph, MBA, BSN, RN, Cherokee Health Systems
Amy Nguyen Howell, MD, MBA, CAPG
Marc Rivo, MD, Population Health Innovations
Julie Schilz, BSN, MBA, Anthem
Pamela Slaven-Lee, DNP, FNP-C, CHSE, The George Washington University School of Nnursing
Lina Walker, PhD, AARP

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 1999

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0062
Measure Title: Comprehensive Diabetes Care: Medical Attention for Nephropathy
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a nephropathy screening test or monitoring test or had evidence of nephropathy during the measurement year.
Developer Rationale: Kidney disease is a major complication of diabetes. The CDC reports that 44% of new kidney failure cases in 2014 were due to diabetes (CDC). In 2013, diabetes led to more than 51,000 cases of kidney failure (Kidney Org). This measure aims to improve the quality of diabetes care through nephrology screenings. Early screenings for people at risk of developing chronic kidney disease can help delay the onset of kidney disease.

Centers for Disease Control and Prevention. National Chronic Kidney Disease Fact Sheet, 2017. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2017.

National Kidney Foundation. Diabetes and Chronic Kidney Disease. 2016.
https://www.kidney.org/news/newsroom/factsheets/Diabetes-And-CKD

Numerator Statement: Patients receiving a nephropathy screening or monitoring test or having evidence of nephropathy during the measurement year
Denominator Statement: Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Exclude patients who did not have a diagnosis of diabetes, in any setting, AND who had a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year
-Exclude patients 65 and older with an advanced illness condition and frailty

Measure Type: Process
Data Source: Claims, Electronic Health Data, Other, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- The developer briefly described the link between nephrology screening or evidence of nephrology and the patient's health outcomes in reducing/improvement in diabetes complications and quality of life.
- The developer provided an updated clinical guideline from the American Diabetes Association (ADA) (2018) including recommendations for the following:
    - Screening- At least once a year, assess urinary (e.g., spot urinary albumin-to-creatinine ratio) and estimated glomerular filtration rate in patients with type 1 diabetes with duration of ≥5 years, in all patients with type 2 diabetes, and in all patients with comorbid hypertension. **B grading**
    - The Level of Evidence grading was A, B, and E for the recommendations provided by developer. A level recommendations used supportive evidence from well-conducted, generalizable, randomized controlled trials. B level recommendations used supportive evidence from a well conducted cohort studies including evidence from well-conducted prospective cohort study or registry; evidence from a well-conducted meta-analysis of cohort studies. E level recommendations used expert consensus or clinical experience.
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guidelines, as this information was not available in the guidelines.
- The developer provided a clinical guideline from the American Geriatrics Society (AGS) (2013) including recommendation for the following:
    - A test for the presence of albuminuria should be performed in individuals at diagnosis of type 2 DM. After the initial screening and in the absence of previously demonstrated macro- or microalbuminuria, a test for the presence of microalbuminuria should be performed annually. **(IIIA)**
    - Quality of Evidence-Level III (definition): Evidence from respected authorities based on clinical experience, descriptive studies, or reports of expert committees
    - Strength of Evidence-A (definition): Good evidence to support the use of a recommendation; clinicians should do this all the time
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guideline, as this information was not available.
- The developer provided a clinical guideline from the American Association of Clinical Endocrinologists (2015) including recommendation for the following:
    - Beginning 5 years after diagnosis in patients with T1D (if diagnosed before age 30) or at diagnosis in patients with T2D and those with T1D diagnosed after age 30, annual assessment of serum creatinine to determine the estimated glomerular filtration rate (eGFR) and urine albumin excretion rate (AER) should be performed to identify, stage, and monitor progression of diabetic nephropathy **(Grade C; Best EL 3)**. Patients with nephropathy should be counseled regarding the need for optimal glycemic control, blood pressure control, dyslipidemia control, and smoking cessation (Grade B; Best EL 2). In addition, they should have routine monitoring of albuminuria, kidney function electrolytes, and lipids (Grade B; Best EL

2). Associated conditions such as anemia and bone and mineral disorders should be assessed as kidney function declines (Grade D; Best EL 4). Referral to a nephrologist is recommended well before the need for renal replacement therapy (Grade D; Best EL 4).
    - The developer did not summarize the Quality, Quantity, and Consistency of the body of evidence associated with the guideline, as this information was not available.
- The developer also provided an additional guideline from the American Association of Clinical Endocrinologists (2011) which provide screening tests recommended by the guideline include microalbumin and serum creatinine. Quality, Quantity, and Consistency of guideline provided.
- The developer also provided a systematic review on the cost-effectiveness of interventions to prevent and control diabetes mellitus. No grading available. Quality, Quantity, and Consistency of systematic review provided.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided updated guidelines from the American Diabetes Association (ADA) (2018), updated guidelines from the American Association of Clinical Endocrinologists (AACE/ACE) (2015), and updated guidelines from the American Geriatrics Society (AGS) (2013) which continues to support their measure focus.
- The developer also provided one additional guideline and one systematic review which provide details on Quantity, Quality, Consistency of the measure focus.

Exception to evidence

NA

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *For structure, process, and intermediate outcome measures:*
    - *What is the relationship of this measure to patient outcomes?*
    - *How strong is the evidence for this relationship?*
    - *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

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

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

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

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

Performance Data:

- Developer provided performance data extracted from HEDIS data, stratified by commercial health plan, Medicare, and Medicaid from 2014, 2015, and 2016.
    - Commercial mean performance- 83.0% (2014) to 89.1% (2016)

- Medicare mean performance- 91.5% (2014) to 95.6% (2016)
- Medicaid mean performance-80.9% (2014) to 89.9% (2016)

- Developer also provided performance data for the NCQA's Diabetes Recognition Program (DRP) from 2015, 2016, and 2017. The mean ranged from 90.2% (2015) to 92.6% (2017)
- Developer provided performance data also from the 2015 PQRS reporting year with a mean of 81.8%.

Disparities

- Developer did not provide disparities data from the measure. However cited CDC data from 2008 that reported incidence of end stage renal disease (ESRD) by race/ethnicity, age, and sex.
  - In 2008, black males and black women had highest incidence of ESRD (461.7 and 304.9 respectively, per 100,000). Hispanic males and Hispanic females followed (271.9 and 205.8 respectively, per 100,000). White men and women had lowest incidence rates (170.7 and 131.5 respectively, per 100,000)
  - In 2008, incidence rates for ESRD were similar among adult 64-74; and older than 75

Questions for the Committee:

- *Specific questions on information provided for gap in care.*
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

Evidence

- The intent of the measure is increase screening for nephropathy in diabetes. The evidence presented is sufficient and related to the desired outcome.
- There is a major limitation here that needs to be called out however: Treatment with ACE inhibitor/ARB is used to qualify as screening having occured. With this being used as such, there may be no incentive to screen/continue to screen patients for progression of disease and intensification or treatment to decrease the risk of progression.
- Also for consideration here is the statement in the ADA 2018 Standards of Care supporting the use of SGLT-2 inhibitors due to evidence suggesting potential clinical benefit in diabetic nephropathy. As further evidence emerges, consideration should be given to including this class of medications as well.
- No additional evidence or studies other than those provided by the Developer. Evidence provided included the American Diabetes Association, American Society of Geriatrics, American Association of Clinical Endocrinologists, and the American Association of Clinical Endocrinologists (2015).
- Diabetic patients are susceptible to diabetic kidney disease. Early detection o of albuminuria, reduced GFR, elevated Creatinine permits optimization of care which may slow or prevent further deterioration of renal function. The developers have updated guidelines using 2018 ADA guidelines, Moderately strong rating for evidence.
- No concerns; developer provided updated evidence from several sources; no need for repeat discussion and vote
- Relationship to pt outcomes: see rationale
- Strength of evidence: moderate
- Evidence applicable to process of care being measured: Yes
- This is a process measure of the percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a nephropathy screening test or monitoring test or had evidence of nephropathy during the measurement year. The included patients had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year. Those diagnosed with gestational diabetes, steroid-induced diabetes, or patients age 65 and older with an advanced illness and frailty were excluded. The American Diabetes Association guideline of 2018 suggests that at least once a year, urine tests (e.g., spot urinary albumin-to-creatinine ratio) and estimated glomerular filtration rate be performed in patients with type 1 diabetes with duration of ≥5 years, in all patients with type 2 diabetes, and in all patients with comorbid hypertension. The American Geriatrics Society guideline for improving diabetes control of 11/13 suggests a test for the presence of albuminuria should be performed in individuals at diagnosis of type 2 DM. Thereafter, a test

for the presence of microalbuminuria should be performed annually in the absence of previously demonstrated macro- or microalbuminuria. The American Association of Clinical Endocrinologists guideline of 2015 recommends an annual assessment of serum creatinine beginning 5 years after diagnosis in patients with type 1 DM (if diagnosed before age 30) or at diagnosis in patients with type 2 DM and those with type 1 DM diagnosed after age 30. An AACE guideline from 2011 also recommended microalbumin as a screening test.

- The development of micro albuminuria is an indicator of potential for progression of diabetic nephropathy. Diabetes is a leading cause of end stage kidney disease. Assessing for the development of diabetic nephropathy can potentially reduce or delay the progression of this complication. The 2018 ADA diabetes treatment guidelines are referenced.
- As the evidence is update and directionally the same and stronger, I don't believe we need to vote on the evidence.
- The measure is very important to patient outcomes. The measure is supported by national guidelines that are evidence based.

Performance Gap

- Data on peformance is provided. It demonstrates improvement recent improvement in performance. Although performance is relatively high, there is still opportunity for improvement. No data on performance by subgroups is presented, however the disproportionate burden of diabetic kidney disease by race/ethnicity if provided.
- Performance data was provided for HEDIS commercial, Medicare and Medicaid, PQRS and the NCQA Diabetes Recognition Program. All results showed some room for improvement.
- CDC disparity data was provided which identified gaps in care for Americans of African descent and Hispanic Americans followed by white Americans."
- Population subgroup was not provided but data from CDC show that ESRD incidence is highest in AA men> AA women> Hispanic men> Hispanic women> non-Hispanic whites men> non-Hispanic white women. How does it demonstrate disparities in the care? Disparities are seen by insurance coverage groups. Medicare beneficiaries have the highest rate of yearly testing for nephropathy.
- According to CDC statistics there are significant disparities in the incidence rate of ESRD among ethnic groups and males and females. For AA males incidence rate of ESRD is 3.5 x more than non-Hispanic white females.
- There is need for improvement regarding performance of this measure for all individuals with Diabetes. Early detection - slows progress of diabetic kidney disease.
- Comment to the current landscape on disparities: Given the current awareness of the role of social determinants of health – it is hard to imagine a system demonstrating quality would be unable to provide this level of data analysis. Most systems collect this data – with this kind of large reporting system, the influence could be great. Also – there are disparity data available to show the need for this kind of stratification – zip codes are usually available data which can support disparity analysis. If certain systems choose to serve populations who struggle in inappropriately designed and fractured systems and then report poorer performance will they be penalized if this measure is used in reimbursement systems?
- HEDIS data from 2014-2016, data from NCQA's Diabetes Recognition Program from 2015-2017, and PQRS data from 2015 are provided. Comprehensive diabetes care remains a HEDIS measure in 2018. CDC studies from 2008 compare the percentages of adult diabetes patients with end stage renal disease by age, sex, and race/ethnicity.
- There has been steady improvement by payers reporting and the Insey QA diabetes recognition program for providers over the preceding three years.
- Information on disparities for this measure are lacking. It is known that African-American males have a much higher incidence of advanced renal disease Amongst individuals with diabetes."
- There is substantial room for improvement as nationally reported performance rates vary from 80 to 91 %. There is no disparity data reported by NCQA but they cited CDC information. Black patients have very high rates of ESRD with DM, Hispanics have lower incidences and whites have the lowest.

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.

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

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: Primary Care and Chronic Illness project team staff

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

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.)?*
- *The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

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

Reliability
- No concerns here.
- Codes and data elements are clearly defined in the code/value sets. The report is able to be consistently reported.
- Data elements are clearly defined.
- Concur with the analysis of the staff evaluator.
- No concerns, developer did empirical reliability testing; no need for discussion and vote
- Rating: high
- Reliability of HEDIS and the Diabetes Recognition Program measures are discussed.
- No concerns
- I have no concerns about reliability of the measure. Further discussion and voting are unwarranted.

Reliability Testing

- No concerns here
- No concerns about the reliability of the measure.
- No concerns regarding reliability testing
- Concur with the analysis of the staff evaluator.
- no concerns
- No
- No concerns
- I have no concerns about reliability of the measure. Further discussion and voting are unwarranted.

Validity Testing

- The only potential threat to validity here is the use of ACE inhibitor/ARB treatment to qualify as meeting the screening criteria. These classes of drugs may be used to treat hypertension in the absence of diabetic kidney disease or may be used in suboptimal doses. The clinical guidelines cited in the evidence portion of the submission recommend screening without any statements related to use of ACE inhibitors/ARBs.
- No concerns with the validity testing of the measure. Differences are shown in HEDIS results between populations such as Medicare, Medicaid and commercial populations.
- No concerns regarding validity testing. No threats to validity
- Concur with the analysis of the staff evaluator.
- No concerns; no testing of exclusions done however the exclusions seem supported by the evidence and testing of distortion by exclusions (frequency of occurrence, variability of exclusion by providers, sensitivity analysis with and without the exclusions) does not seem warranted;
- Validity of HEDIS measures and data from the Diabetes Recognition Program are discussed.
- No concerns
- I have no concerns about the validity of the measure. Further discussion and voting are unwarranted.

Other threats

- No concerns
- New exclusion in the measure includes "frailty" which is more challenging. No concerns with validity and no real threats to validity.
- No threats to validity related to exclusions and risk adjustment
- Concur with the analysis of the staff evaluator.
- Construct validity testing at level of health plan and provider supports use of this as quality measure
- No concerns
- I have no concerns about the validity of the measure. Further discussion and voting are unwarranted.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

Data Specifications and Elements

- The measure is constructed using multiple data sources (administrative data, electronic clinical data, and paper records)
- Some data elements are in defined fields in electronic sources
- Developer shared no difficulties on the use of this measure in HEDIS or NCQA's Diabetes Recognition Program.
- This is not an eMeasure

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Feasibility

- No concerns
- The measure is feasible as it uses both administrative and chart data. More human resource intensive to collect and more costly. E-measure would make it easier once improvements occur in the availability of electronic health record data for purposes of measure reporting.
- Data elements are routinely generated and used during care delivery. With wider use of electronic health records data generation will be more efficient
- Comment on eMeasure responses: There is a super majority of providers using EMR/EHRs – the response given seems to be out of sync with where the systems of care actually are - utilizing electronic medical records, and those that aren't, should be for many reasons, patient safety being a primary one. There is no described path to an eMeasure either.
- No concerns, already collecting data
- The HEDIS Audit process is described.
- No concerns. Feasible from claims data or data from the electronic health record
- Feasibility is high as the test is easily performed, and multiple, easily accessed electronic data sources capture the information.

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

- HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website.
- STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.
- CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).
- INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States.

- ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care.
- QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance.

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

This measure uses the following methods to obtain input: including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.

NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System.

Feedback on the measure by those being measured or others

Questions received through NCQA's Policy Clarification Support system have generally centered around clarification on types of lab tests that are considered screening or monitoring for nephropathy such as creatine/glomerular filtration and urinalysis or documentation of history of mico albuminuria or if patient must be on an ACE/ARB the entire measurement year to be counted in the measure. In response, the developer has provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

Additional Feedback:

The developer/steward did not provide any further feedback.

***Questions for the Committee**:*

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

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

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

4b. Usability 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

- From 2014 to 2016, performance rates for this measure have increased for all product lines (Commercial, Medicaid, and Medicare). Of the plans, the highest performance continues to be seen in the Medicare population. In 2016, Medicare plans had a performance rate of 97 percent while Commercial and Medicaid has around 90 percent (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). These data are nationally representative.

- The developer states that performance rates have slightly gone up, despite a decrease in the number of reporting physicians seeking recognition in the NCQA's Diabetes Recognition Program since 2015.

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

- Per developer, there were no identified unexpected findings (positive or negative) during testing or since implementation of this measure.

Potential harms

- The developer briefly cited in evidence form the following:
  - The harms associated with the screening and treatment of nephropathy stem from adverse effects that are associated with pharmacotherapy and other treatment options (dialysis and kidney transplant). One study suggested higher risks to patients when using combined medication therapies as opposed to monotherapy (Halimi et al., 2009).

Additional Feedback:
NA

***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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Use

- This measure is used in multiple accountability programs publicly reported.
- Part of Health Plan report cards, NCQA State of Health Care report, CMS QPP, California P4P, accreditation of ACO and Health Plan by NCQA, Quality Compass
- Measure is used for reporting in NCQA's Diabetes Recognition Program and in HEDIS (including Quality Compass).
- The measure is being publicly reported and feedback is being given to the eligible providers
- "How is the value communicated to the patient – is it only used by the system?
- Overall Feedback Responses: How are patients and consumers meaningfully engaged in the development and implementation of the measure? It is unclear from the responses where and how this occurred. Ultimately patients are the "measured" entity.
- No concerns
- HEDIS data are published in numerous publications and many types of providers reference HEDIS data.
- This measure is being publicly reported for use by health plan report cards, CMS quality payment program and accountable care organization accreditation. Those being measured are given performance results are data.
- easure performance is reported via numerous national channels. It is also a measure utilized for ACO accreditation.
- Multi-stakeholder advisory panels and public commenting as well as questions for clarification.

Usability

- Through HEDIS reporting, measure results are used to compare health plans. In the Diabetes Recognition Program results are used to compare providers. No concerns with usability. No identified harms in reporting measure results.
- Measure needs to be applied more widely, to every patient with diabetes. No harms identified.
- There are many great examples of how these outcomes are communicated to providers but fewer on how these

data are communicated back to patients. One would expect equally robust outreach to patients – are any of the conferences patient-centered conferences or are they provider facing?
- No concerns, except potential adverse events associated with treatment;
- From 2014 to 2016, performance rates from HEDIS data show increases for this measure. The developer states that performance rates have slightly gone up in the NCQA Diabetes Recognition Program, despite a decrease in the number of reporting physicians since 2015. There were no identified unexpected findings (positive or negative) during testing or since implementation of this measure. The harms associated with the screening and treatment of nephropathy stem from adverse effects that are associated with pharmacotherapy and other treatment options (dialysis and kidney transplants).
- Benefits in this measure certainly outweigh any harms of measurement. One concern, however, as this measure is written would be that individual with stage III chronic kidney disease who has microalbuminuria may not have ongoing albuminuria monitoring. If this individual developed macroalbuminuria, consultation with nephrology is highly recommended(KDOQI), yet as this measure is written, there could be a delay in identification of the macroalbuminuria and thus delay in consultation and treatment.
- "ince the last endorsement, performance rates have increased across all plan types.
- There are no harms to screening
- Potential harms of ACE/ARB therapy include adverse reactions or interactions with other classes of medications resulting in adverse events.

Criterion 5: Related and Competing Measures

Related or competing measures

- Developer did not identify any related or competing measures.

Harmonization

NA

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments were received.

Measure Number: Comprehensive Diabetes Care: Medical Attention for Diabetic Nephropathy

Measure Title: NQF# 0062

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document
   NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
   TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

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

The measure's reliability per beta-binomial model for the physician level is 0.90. The measure's reliability per beta-binomial model for the health plan (commercial, medicare, Medicaid) level are 1.00, 0.97, and 0.97. These results indicates the measure has high reliability.

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

Beta-binomial calculation was used to test measure score reliability.

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

Reliability was assessed from physician/practice data from the NCQA Diabetes Recognition Program that included 3676 physicians for the time frame of 2010-2012.

Reliability was assessed from HEDIS dat that included 416 commerical health plans, 500 Medicare health plans, and 194 Medicaid health plans.

4. 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.*

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level) -** What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

The measure's reliability per beta-binomial model for the physician level is 0.90 and for the health plan level is 0.97-1.00. These results indicates the measure has high reliability, meaning that differences in physicin/health plan performance reflect true differences in quality as opposed to measurement error or noise.

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

   REFERENCE: Testing attachment, section 2a2.2
   ***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☐Yes (go to Question #8)

   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

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

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

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

There was no testing of the exclusions done.

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

REFERENCE: Testing attachment, section 2b2.

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

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

There was no testing of the exclusions done.

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

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

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.
☐Yes (please explain below then go to Question #15)
☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)
No missing data and "measure is collected with a complete sample" per developer.

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)
Developer did construct validity testing.

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

The developer tested for construct validity at the health plan level by exploring whether the measure was correlated with other similar measures of quality hypothesized to be related.

The developer tested for construct validity at the physician level by exploring whether the measure was correlated with other similar measures of quality in NCQA's Diabetes Recognition Program hypothesized to be related.

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    - ☐High (go to Question #21)
    - ☒Moderate (go to Question #21)
    - ☐Low (please explain below then go to Question #21)
    - ☐Insufficient (go to Question #21)

At health plan level, the correlations are moderate to strong and statistically significant. These results confirmed the hypothesis that the diabetes measures are correlated with each other.

At physician level, the correlations are moderate to weak. Per developer, overall these correlation results suggest that the physician level measure has sufficient validity.

21. Was validity testing conducted with <u>patient-level data elements</u>?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Prior validity studies of the same data elements may be submitted*
    - ☐Yes (go to Question #22)
    - ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

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.*
    ***REFERENCE:** Testing attachment, section 2b1.*
    ***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*
    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    - ☐Yes (go to Question #23)
    - ☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
    - ☐Moderate (skip Questions #24-25 and go to Question #26)
    - ☐Low (please explain below, skip Questions #24-25 and go to Question #26)
    - ☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has <u>not</u> been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was <u>face validity</u> systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
    NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
    REFERENCE: Testing attachment, section 2b1.
    ***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*
    ☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)
    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)
    ☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.
    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)
    ☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]
    ☐Insufficient (if insufficient, please explain below) [NOTE: For 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—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?
    REFERENCE: Testing attachment, section 2c
    ***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*
    ☐High
    ☐Moderate
    ☐Low (please explain below)
    ☐Insufficient (please explain below)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0062

Measure Title: Comprehensive Diabetes Care: Medical Attention for Nephropathy

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

Date of Submission: 4/9/2018

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: receiving a nephropathy screening test or having evidence of nephropathy during the measurement year.

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Adults with diabetes (type 1 or 2) >>> Nephropathy screening is performed or evidence of nephropathy is documented>>> Screening results are evaluated >>>Results indicative of nephropathy>>>Health provider determines treatment to delay progression of diabetic nephropathy>>>improvement in diabetes complications and quality of life.

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.)

N/A

****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

Table 1. American Diabetes Association (ADA) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Diabetes Association. (2018). Standards of Medical Care in Diabetes – 2018. Diabetes Care 2018; 41(Suppl. 1): S105-S118; doi: 10.2337/dc18-S010 Guideline available from: http://care.diabetesjournals.org/content/41/Supplement_1 2013 Submission American Diabetes Association. (2013). Standards of Medical Care in Diabetes – 2013. Diabetes Care 2013; 36:S1-e4; doi: 10.2337/dc13-S001 Guideline available from: http://care.diabetesjournals.org/content/36/Supplement_1/S11
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the	2018 Submission Pg. S105-106 “Screening

conclusions from the SR.	• At least once a year, assess urinary (e.g., spot urinary albumin-to-creatinine ratio) and estimated glomerular filtration rate in patients with type 1 diabetes with duration of ≥5 years, in all patients with type 2 diabetes, and in all patients with comorbid hypertension. (B) Treatment • In nonpregnant patients with diabetes and hypertension, either an ACE inhibitor or an angiotensin receptor blocker is recommended for those with modestly elevated urinary albumin–to–creatinine ratio (30–299 mg/g creatinine) (B) and is strongly recommended for those with urinary albumin–to creatinine ratio ≥300 mg/g creatinine and/or estimated glomerular filtration rate <60 mL/min/1.73 m2 (A) • Periodically monitor serum creatinine and potassium levels for the development of increased creatinine or changes in potassium when ACE inhibitors, angiotensin receptor blockers, or diuretics are used. (B) • Continued monitoring of urinary albumin–to–creatinine ratio in patients with albuminuria treated with an ACE inhibitor or an angiotensin receptor blocker is reasonable to assess the response to treatment and progression of diabetic kidney disease. (E) • An ACE inhibitor or an angiotensin receptor blocker is not recommended for the primary prevention of diabetic kidney disease in patients with diabetes who have normal blood pressure, normal urinary albumin–to–creatinine ratio (<30 mg/g creatinine), and normal estimated glomerular filtration rate. (B) • When estimated glomerular filtration rate is <60 mL/min/1.73 m2, evaluate and manage potential complications of chronic kidney disease. (E) • Patients should be referred for evaluation for renal replacement treatment if they have an estimated glomerular filtration rate <30 mL/min/1.73 m2. (A) • Promptly refer to a physician experienced in the care of kidney disease for uncertainty about the etiology of kidney disease, difficult management issues, and rapidly progressing kidney disease. (B) 2013 Submission

	Pg. S7-S8 "Screening • Perform an annual test to assess urine albumin excretion in type 1 diabetic patients with diabetes duration of ≥5 years and in all type 2 diabetic patients starting at diagnosis. (B) • Measure serum creatinine at least annually in all adults with diabetes regardless of the degree of urine albumin excretion. The serum creatinine should be used to estimate glomerular filtration rate (GFR) and stage the level of chronic kidney disease (CKD), if present. (E) Treatment • In the treatment of the nonpregnant patient with modestly elevated (30–299 mg/day) (C) or higher levels (≥300 mg/day) of urinary albumin excretion, either ACE inhibitors or ARBs are recommended. (A) • Reduction of protein intake to 0.8–1.0g/kg body wt per day in individuals with diabetes and the earlier stages of CKD and to 0.8 g/kg body wt per day in the later stages of CKD may improve measures of renal function (urine albumin excretion rate, GFR) and is recommended. (C) • When ACE inhibitors, ARBs, or diuretics are used, monitor serum creatinine and potassium levels for the development of increased creatinine or changes in potassium. (E) • Continued monitoring of urine albumin excretion to assess both response to therapy and progression of disease is reasonable. (E) • When eGFR is <60 mL/min/1.73 m2, evaluate and manage potential complications of CKD. (E) • Consider referral to a physician experienced in the care of kidney disease for uncertainty about the etiology of kidney disease, difficult management issues, or advanced kidney disease. (B)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Level of evidence and description: • A: Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted multicenter trial o Evidence from a meta-analysis that incorporated quality ratings in the analysis

	Compelling nonexperimental evidence, i.e., “all or none” rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted trial at one or more institutions o Evidence from a meta-analysis that incorporated quality ratings in the analysis • B: Supportive evidence from well-conducted cohort studies, including: o Evidence from a well-conducted prospective cohort study or registry o Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study • E: Expert consensus or clinical experience <u>2013 Submission</u> Level of Evidence & Description: • A: Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted multicenter trial o Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., “all or none” rule developed by the Centre for Evidence-Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: o Evidence from a well-conducted trial at one or more institutions o Evidence from a meta-analysis that incorporated quality ratings in the analysis • B: Supportive evidence from well-conducted cohort studies,

	including: o Evidence from a well-conducted prospective cohort study or registry o Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study • C Supportive evidence from poorly controlled or uncontrolled studies o Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results o Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) o Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation • E: Expert consensus or clinical experience
Provide all other grades and definitions from the evidence grading system	2018 Submission Level of Evidence & Description: • C Supportive evidence from poorly controlled or uncontrolled studies o Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results o Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) o Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation 2013 Submission No additional grades aside from what is listed above
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above

	2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The ADA does not provide information on the systematic review conducted to support its 2018 or 2013 guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on two other systematic reviews that support the ADA's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 2. American Association of Clinical Endocrinologists (AACE)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission AACE/American College of Endocrinology (ACE). Clinical Practice Guidelines for Developing a Diabetes Mellitus Comprehensive Care Plan-2015. 2015. Endocrine Practice. Vol 21 (Suppl1). URL: https://www.aace.com/files/dm-guidelines-ccp.pdf 2013 Submission AACE. Medical Guidelines for Clinical Practice For Developing A Diabetes Mellitus Comprehensive Care Plan. Endocrine Practice. 2011 Vol 17, Suppl 2: 1-53 **URL:** http://journals.aace.com/doi/abs/10.4158/EP.17.S2.1

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.	2018 Submission Pg. 16 "3.Q9 – Recommendation 28 • Beginning 5 years after diagnosis in patients with T1D (if diagnosed before age 30) or at diagnosis in patients with T2D and those with T1D diagnosed after age 30, annual assessment of serum creatinine to determine the estimated glomerular filtration rate (eGFR) and urine albumin excretion rate (AER) should be performed to identify, stage, and monitor progression of diabetic nephropathy (**Grade C; Best EL 3**). Patients with nephropathy should be counseled regarding the need for optimal glycemic control, blood pressure control, dyslipidemia control, and smoking cessation (**Grade B; Best EL 2**). In addition, they should have routine monitoring of albuminuria, kidney function electrolytes, and lipids (**Grade B; Best EL 2**). Associated conditions such as anemia and bone and mineral disorders should be assessed as kidney function declines (**Grade D; Best EL 4**). Referral to a nephrologist is recommended well before the need for renal replacement therapy (**Grade D; Best EL 4**). 2013 Submission Pg. 11 ***"3.Q10.1. Diabetic Nephropathy*** • **R36.** Beginning 5 years after diagnosis in patients with T1DM and at diagnosis in patients with T2DM, an annual assessment of serum creatinine to estimate the glomerular filtration rate (GFR) and urine albumin excretion should be performed to identify, stage, and monitor progression of diabetic nephropathy (**Grade D; Best EL 4**). Patients with diabetic nephropathy should be counseled regarding the increased need for optimal glycemic control, blood pressure control, dyslipidemia control, and smoking cessation (**Grade A; Best EL 1**). When therapy with angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers is initiated, renal function and serum potassium levels must be closely monitored (**Grade A; Best EL 1**)."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission **Numerical Descriptor (evidence level)** 2 Meta-analysis of nonrandomized prospective or case-controlled trials (MNRCT) 2 Nonrandomized controlled trial (NRCT) 2 Prospective cohort study (PCS) 2 Retrospective case-control study (RCCS)

	3 Cross-sectional study (CSS) 3 Surveillance study (registries, surveys, epidemiologic study, retrospective chart review, mathematical modeling of database) (SS) 3 Consecutive case series (CCS) 3 Single case reports (SCR) 4 No evidence (theory, opinion, consensus, review, or preclinical study) (NE) a Adapted from (1): *Endocr Pract*. 2010;16:270-283. b 1, strong evidence; 2, intermediate evidence; 3, weak evidence; and 4, no evidence. 2013 Submission 1 Meta-analysis of randomized controlled trials (MRCT) 1 Randomized controlled trials (RCT) 4 No evidence (theory, opinion, consensus, review, or preclinical study) (NE) a Adapted from (1): *Endocr Pract*. 2010;16:270-283. b 1, strong evidence; 2, intermediate evidence; 3, weak evidence; and 4, no evidence.
Provide all other grades and definitions from the evidence grading system	2018 Submission 1 Meta-analysis of randomized controlled trials (MRCT) 1 Randomized controlled trials (RCT) a Adapted from (1): *Endocr Pract*. 2010;16:270-283. b 1, strong evidence; 2, intermediate evidence; 3, weak evidence; and 4, no evidence. 2013 Submission 2 Meta-analysis of nonrandomized prospective or case-controlled trials (MNRCT) 2 Nonrandomized controlled trial (NRCT) 2 Prospective cohort study (PCS) 2 Retrospective case-control study (RCCS) 3 Cross-sectional study (CSS) 3 Surveillance study (registries, surveys, epidemiologic study, retrospective chart review, mathematical modeling of database) (SS) 3 Consecutive case series (CCS)

3 Single case reports (SCR)

a Adapted from (1): *Endocr Pract*. 2010;16:270-283.

b 1, strong evidence; 2, intermediate evidence; 3, weak evidence; and 4, no evidence.

Grade assigned to the **recommendation** with definition of the grade

2018 Submission

Grading of Recommendations; How Different Evidence Levels Can Be Mapped to the Same Recommendation Grade

Best evidence level	Subjective factor impact	Two-thirds consensus	Mapping	Recommendation Grade
2	None	Yes	Direct	B
1	Negative	Yes	Adjust down	B
3	Positive	Yes	Adjust up	B
3	None	Yes	Direct	C
2	Negative	Yes	Adjust down	C
4	Positive	Yes	Adjust up	C
4	None	Yes	Direct	D
3	Negative	Yes	Adjust down	D
1,2,3,4	Positive	No	Adjust down	D

Starting with the left column, best evidence levels (BELs), subjective factors, and consensus map to recommendation grades in the right column. When subjective factors have little or no impact ("none"), then the BEL is directly mapped to recommendation grades. When subjective factors have a strong impact, then recommendation grades may be adjusted up ("positive" impact) or down ("negative" impact). If a two-thirds consensus cannot be reached, then the recommendation grade is D. NA, not applicable (regardless of the presence or absence of strong subjective factors, the absence of a two-thirds consensus mandates a recommendation grade D).

Reprinted from reference 1: *Endocr Pract*. 2010;16:270-283.

2013 Submission

Grading of Recommendations; How Different Evidence Levels Can Be Mapped to the Same Recommendation Grade

	Best evidence level / Subjective factor impact / Two-thirds consensus / Mapping / Recommendation grade 1 / None / Yes / Direct / A 2 / Positive / Yes / Adjust up / A 4 / None / Yes / Direct / D 3 / Negative / Yes / Adjust down / D 1, 2, 3, 4 / NA / No / Adjust down / D Starting with the left column, best evidence levels (BELs), subjective factors, and consensus map to recommendation grades in the right column. When subjective factors have little or no impact ("none"), then the BEL is directly mapped to recommendation grades. When subjective factors have a strong impact, then recommendation grades may be adjusted up ("positive" impact) or down ("negative" impact). If a two-thirds consensus cannot be reached, then the recommendation grade is D. NA, not applicable (regardless of the presence or absence of strong subjective factors, the absence of a two-thirds consensus mandates a recommendation grade D).
Provide all other grades and definitions from the recommendation grading system	2018 Submission Best evidence level / Subjective factor impact / Two-thirds consensus / Mapping / Recommendation Grade 1 / None / Yes / Direct / A 2 / Positive / Yes / Adjust up / A 2013 Submission Best evidence level / Subjective factor impact / Two-thirds consensus / Mapping / Recommendation Grade 2 / None / Yes / Direct / B 1 / Negative / Yes / Adjust down / B 3 / Positive / Yes / Adjust up / B 3 / None / Yes / Direct / C 2 / Negative / Yes / Adjust down / C 4 / Positive / Yes / Adjust up / C

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AACE guideline evidence review is listed in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See Table 4 below

Table 3. American Geriatrics Society (AGS) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission American Geriatrics Society (AGS). 2013. Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. American Geriatrics Society Panel on the Care for Older Adults with Diabetes Mellitus. Journal of American Geriatric Society. 2013 November; 61 (11): 2020-2026. Doi:10.1111/jgs.12514 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064258/pdf/nihms583558.pdf 2013 Submission American Geriatrics Society (AGS). 2003. Guidelines for Improving the Care of the Older Person with Diabetes Mellitus. California Healthcare Foundation/American Geriatrics Society Panel on Improving Care for Elders

	with Diabetes. American Geriatrics Society. May 2013; 51, Suppl 5, JAGS URL:
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.	2018 Submission "A test for the presence of albuminuria should be performed in individuals at diagnosis of type 2 DM. After the initial screening and in the absence of previously demonstrated macro- or microalbuminuria, a test for the presence of microalbuminuria should be performed annually. (IIIA) 2013 Submission Pg. 272 "A test for the presence of microalbumin should be performed at diagnosis in patients with type 2 DM. After the initial screening and in the absence of previously demonstrated macro- or microalbuminuria, a test for the presence of microalbumin should be performed annually. (IIIA)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission Quality of Evidence • Level III: Evidence from respected authorities based on clinical experience, descriptive studies, or reports of expert comittees Strength of Evidence • A: Good evidence to support the use of a recommendation; clinicians should do this all the time 2013 Submission Same as above
Provide all other grades and definitions from the evidence grading system	2018 Submission Quality of Evidence • Level I: Evidence from at least one properly randomized controlled trial • Level II: Evidence from at least one well-designed clinical trial without randomization, from cohort or case-controlled analytical studies, from multiple time-series, or from dramatic results in uncontrolled experiments Strength of Evidence • B: Moderate evidence to support the use of a recommendation clinicians "should do this most of the time" • C: Poor evidence to support or to reject the use of a recommendation; clinicians may or may not follow the recommendation • D: Moderate evidence against the use of a recommendation; clinicians

	should not do this • E: Good evidence against the use of a recommendation; clinicians should not do this 2013 Submission Same as above
Grade assigned to the **recommendation** with definition of the grade	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Provide all other grades and definitions from the recommendation grading system	2018 Submission No additional grading was provided for the recommendations aside from what is described above 2013 Submission No additional grading was provided for the recommendations aside from what is described above
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The AGS does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the AGS systematic review, we provide information on two other systematic reviews that support the AGS's recommendations in Table 4.
Estimates of benefit and consistency across studies	See Table 4 below
What harms were identified?	See Table 4 below
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See Table 4 below

Table 4. Additional Systematic Reviews

Citations	AACE Diabetes Care Plan Guidelines. Endocrine Practice.	Li R, Zhang P, Barker LE, Chowdhury FM, Zhang X. Cost-effectiveness of interventions to

	2011. Vol 17, Suppl 2: 1-53 **URL:** http://journals.aace.com/doi/abs/10.4158/EP.17.S2.1	prevent and control diabetes mellitus: a systematic review. Diabetes Care. 2010. 33(8):1872-1894. **URL:** http://care.diabetesjournals.org/content/33/8/1872.full.pdf+html
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	This measure assesses whether diabetic patients were screened for nephropathy if they did not already have evidence of nephropathy during the measurement year. The measure is based on clinical guidelines. Evidence provides support for the timing of screenings, specific screening tests, and treatment based on screening results. Screening tests recommended by the guideline include microalbumin and serum creatinine. Treatment recommendations from the guidelines include medications, counseling, nephrologist referral, close monitoring of urine albumin excretions, and close monitoring of nephropathy progression.	
Grade assigned for the quality of the quoted evidence with definition of the grade	Numerical descriptor (evidence level) / Semantic descriptor (reference methodology) 1 — Randomized controlled trials (RCT) 2 — Meta-analysis of nonrandomized prospective or case-controlled trials (MNRCT) 2 — Nonrandomized controlled trial (NRCT) 2 — Prospective cohort study (PCS) 3 — Cross-sectional study	No grading provided

	(CSS) 3 Surveillance study (registries, surveys, epidemiologic study, retrospective chart review, mathematical modeling of database) (SS) 4 No evidence (theory, opinion, consensus, review, or preclinical study) (NE) 1=strong evidence; 2=intermediate evidence; 3=weak evidence; and 4=no evidence.	
Provide all other grades and associated definitions of the evidence in the grading system	1 Meta-analysis of randomized controlled trials (MRCT) 2 Retrospective case-control study (RCCS) 3 Single case reports (SCR) 3 Consecutive case series (CCS)	N/A
What is the time period covered by the body of evidence?	1993-2008	1993-2007
Quantity and Quality of Body of Evidence	<u>Screening</u> Measurement of albumin to creatinine ratio: Clinical Practice Guideline No Evidence Use of glomerular filtration rate (GFR) in screening for nephropathy:	Seventeen studies for interventions end stage renal disease or nephropathy were identified. The interventions included screenings for microalbuminuria and treatment options to delay the progression of nephropathy. The studies included

	1 Cross-sectional study Estimation of GFR: 1 surveillance study <u>Treatment</u> Medication treatment to prevent onset or delay progression of diabetic nephropathy: 4 randomized controlled trials, 1 Prospective cohort study, 2 Review/no evidence Normalization of albumin excretion to decrease nephropathy progression: 2 randomized controlled trials Restricting protein intake in patients nephropathy: 1 meta-analysis of nonrandomized prospective or case-controlled trials Referral of stage 4 chronic kidney disease patients to nephrologist: opinion/no evidence	RCTs, cohort studies, observational studies, and clinical trials.
What is the overall quality of evidence <u>across studies</u> in the body of evidence?	The overall quality of evidence for the measure focus is high. Guidelines supporting the measure include recommendations for the screening and treatment of nephropathy. Evidence for treatment options to prevent nephropathy onset and delay the progression of nephropathy have the strongest evidence with the most RCTs. The evidence supporting screenings for nephropathy is weaker in comparison to the nephropathy treatment evidence. This evidence includes clinical trials, cross sectional studies, surveillance studies, and large cohorts studies as opposed to RCTs. Evidence for nephropathy screenings also include literature reviews. Despite this weaker evidence for nephropathy screenings, the linkage to improved nephropathy outcomes through screening is high. Regular nephropathy screenings offer the opportunity for early detection of diabetic nephropathy and early treatment to delay progression of the disease.	
Estimates of benefit and consistency across studies in body of evidence – what	The evidence supporting this measure can be categorized into evidence for nephropathy screening and evidence for nephropathy treatments. Screening is a crucial step in delaying the onset or progression of nephropathy in diabetics. The results from one study cited that the average life expectancy increases from four to 14 years with nephropathy screening and interventions (Borch-Johnson, 1993). In addition, the study cited a	

are the estimates of benefits?	decrease in the need for dialysis and kidney transplants by 21% to 63% (Borch-Johnson, 1993). The onset of nephropathy is can also be delayed by six to 24 years and therefore, reduces the mortality rates of deaths due to nephropathy (Borch-Johnsen, 1993). Another treatment method identified by the guidelines supporting this measure includes referral to a nephrologist. An important aspect of referral includes timeliness. Data suggests that the early referral to a nephrologist can improve mortality rates and lifespan of patients on dialysis. Patients that begin treatment with a nephrologist over a year before starting dialysis live longer lives, on average, than patients that were referred within four months of starting dialysis. Screening for nephropathy is a necessary component of determining the stage of kidney disease. Therefore, the benefit of regular screenings will lead to earlier specialized treatment and improved outcomes for diabetic nephropathy. Borch-Johnsen K, Wenzel H, Viberti GC, Mogensen CE. Is screening and intervention for microalbuminuria worthwhile in patients with insulin dependent diabetes? BMJ. 1993; 306: 1722-1725.
What harms were studied and how do they affect the net benefit (benefits over harms)?	The harms associated with the screening and treatment of nephropathy stem from adverse effects that are associated with pharmacotherapy and other treatment options (dialysis and kidney transplant). One study suggested higher risks to patients when using combined medication therapies as opposed to monotherapy (Halimi et al., 2009). Halimi JM, Asmar R, Ribstein J. Optimal nephroprotection: Use, misuse and misconceptions about blockade of the renin-angiotensin system. Lessons from the ONTARGET and other recent trials. Diabetes Metab. 2009; 35:425-430.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous studies have been conducted since the systematic reviews we cite in this table, none of which change the conclusion that medical attention for nephropathy for individuals with diabetes is appropriate.

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.

N/A

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

N/A

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

N/A

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0062
Corresponding Measures:
De.2. **Measure Title:** Comprehensive Diabetes Care: Medical Attention for Nephropathy
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18-75 years of age with diabetes (type 1 and type 2) who received a nephropathy screening test or monitoring test or had evidence of nephropathy during the measurement year.
1b.1. **Developer Rationale:** Kidney disease is a major complication of diabetes. The CDC reports that 44% of new kidney failure cases in 2014 were due to diabetes (CDC). In 2013, diabetes led to more than 51,000 cases of kidney failure (Kidney Org). This measure aims to improve the quality of diabetes care through nephrology screenings. Early screenings for people at risk of developing chronic kidney disease can help delay the onset of kidney disease.
Centers for Disease Control and Prevention. National Chronic Kidney Disease Fact Sheet, 2017. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2017.

National Kidney Foundation. Diabetes and Chronic Kidney Disease. 2016. https://www.kidney.org/news/newsroom/factsheets/Diabetes-And-CKD

S.4. **Numerator Statement:** Patients receiving a nephropathy screening or monitoring test or having evidence of nephropathy during the measurement year
S.6. **Denominator Statement:** Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.
S.8. **Denominator Exclusions:** Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Exclude patients who did not have a diagnosis of diabetes, in any setting, AND who had a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year
-Exclude patients 65 and older with an advanced illness condition and frailty

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data, Other, Paper Medical Records
S.20. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Sep 02, 2014

IF this measure is included in a composite, NQF Composite#/title:
0731:Comprehensive Diabetes Care

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

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

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and

improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_0062_Nephropathy_7.1.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.
Yes

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.
Kidney disease is a major complication of diabetes. The CDC reports that 44% of new kidney failure cases in 2014 were due to diabetes (CDC). In 2013, diabetes led to more than 51,000 cases of kidney failure (Kidney Org). This measure aims to improve the quality of diabetes care through nephrology screenings. Early screenings for people at risk of developing chronic kidney disease can help delay the onset of kidney disease.

Centers for Disease Control and Prevention. National Chronic Kidney Disease Fact Sheet, 2017. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2017.

National Kidney Foundation. Diabetes and Chronic Kidney Disease. 2016. https://www.kidney.org/news/newsroom/factsheets/Diabetes-And-CKD

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.*
HEDIS
The following data are extracted from HEDIS data collection reflecting the most recent years of reporting for this measure. Performance data is summarized at the health plan level and summarized by number of plans reporting, mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year and product line (i.e. commercial, Medicare, and Medicaid).

Comprehensive Diabetes Care: Medical Attention for Nephropathy
*Higher score= better performance
N= Number of plans reporting

Commercial Rate
YEAR | N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2014| 404| 83.0%| 5.3%| 68.5%| 76.0%| 79.4%| 83.0%| 86.7%| 89.6%| 97.8%
2015| 419| 88.9%| 3.5%| 74.1%| 84.6%| 87.3%| 89.1%| 91.2%| 93.1%| 97.9%
2016| 412| 89.1%| 3.0%| 76.9%| 85.5%| 87.4%| 89.3%| 91.0%| 92.6%| 99.6%

Medicaid Rate
YEAR | N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2014| 220| 80.9%| 6.2%| 57.0%| 73.8%| 77.9%| 81.8%| 84.9%| 87.7%| 100.0%
2015| 261| 90.0%| 3.2%| 74.1%| 86.1%| 88.6%| 90.5%| 92.0%| 93.5%| 97.2%
2016| 271| 89.9%| 3.5%| 69.6%| 86.7%| 88.6%| 90.3%| 91.7%| 93.3%| 99.8%

Medicare Rate
YEAR | N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX

2014| 475| 91.5%| 3.8%| 66.3%| 87.2%| 89.6%| 91.7%| 94.0%| 95.6%| 100.0%
2015| 461| 95.3%| 2.6%| 72.9%| 92.5%| 94.2%| 95.6%| 96.9%| 98.1%| 100.0%
2016| 473| 95.6%| 2.4%| 79.2%| 92.7%| 94.2%| 95.8%| 97.3%| 98.2%| 100.0%

This measure is used NCQA's Diabetes Recognition Program (DRP) that assesses clinician performance on key quality measures that are based on national evidence based guidelines in diabetes care (see full description of program in 4a1.1). Below is performance data for this measure in the program.

Diabetes Recognition Program
YEAR|N|MEAN|ST DEV|MIN|10TH|25TH|50TH|75TH|90TH|MAX
2015|4989|90.2%|14.1%|0.0%|76.5%|88.0%|94.8%|97.5%|100.00%|100.00%
2016|4704|91.9%|11.5%|0.00%|80.0%|90.0%|96.0%|98.8%|100.00%|100.00%
2017|3771|92.6%|11.5%|0.00%|83.1%|92.0%|96.0%|100.0%|100.00%|100.00%

PQRS
The following PQRS performance data includes claims, registry, measures group, GPRO Web Interface/ACO, QCDR data for services performed from in 2015.

Mean: 81.8%
St dev: 16.9%

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.
N/A

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.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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
2008 CDC data reports the incidence of end stage renal disease (ESRD) among diabetics to be highest in black males (461.7 per 100,000) and black women (304.9). The incidence rates for Hispanic men and women in 2008 were 271.8 and 205.8, respectively. White men and women with diabetes had the lowest incidence rates for ESRD at 170.7 and 131.5, respectively. The incidence rates for ESRD were similar among adults in ages 65-74 (319.7) and older than 75 (317.7). The incidence rates were reported per 100,000 diabetic population (CDC, 2012).

Centers for Disease Control and Prevention. 2012. CDC's Diabetes Program-Data and Trends-End Stage Renal Disease-Age-Adjusted Incidence of End Stage Renal Disease Related to Diabetes Mellitus by Race/Ethnicity and Sex.

Centers for Disease Control and Prevention. 2012. CDC's Diabetes Program-Data and Trends-End Stage Renal Disease-Age-Adjusted

Incidence of End Stage Renal Disease Related to Diabetes Mellitus by Age.

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):*
Endocrine, Endocrine : Diabetes, Renal, Renal : Chronic Kidney Disease (CKD)

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):*
Populations at Risk

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.)*
N/A

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0062_CDC_Nephropathy_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

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

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Added another optional exclusion which is to exclude patients 65 and older with an advanced illness condition and frailty. This was added because quality measures that were intended for the general population may not be clinically appropriate or priority for individuals with advanced illness.

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

measure.
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients receiving a nephropathy screening or monitoring test or having evidence of nephropathy during the measurement year

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Time period for data: a measurement year (12 months)

ADMINISTRATIVE CLAIMS: Due to the extensive volume of codes associated with identifying numerator events for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD: At a minimum, documentation in the medical record must include a note indicating the date when the nephropathy screening or monitoring test was performed or nephropathy evidence documented. The patient is numerator compliant if the nephropathy screening was performed or nephropathy evidence is documented. The patient is not numerator compliant if nephropathy screening and result are missing or if nephropathy evidence is not documented. Ranges and thresholds do not meet criteria for this measure.

Any of the following meet criteria for a nephropathy screening or monitoring test of evidence of nephropathy:

-A urine test for albumin or protein (At a minimum, documentation must include a note indicating the date when a urine test was performed, and the result or finding. Documentation includes: 24-hour urine for albumin or protein, Timed urine for albumin or protein., Spot urine (e.g., urine dipstick or test strip) for albumin or protein, Urine for albumin/creatinine ratio, 24-hour urine for total protein, random urine for protein/creatinine ratio.)

-Documentation of a visit to a nephrologist.

-Documentation of a renal transplant.

-Documentation of medical attention for any of the following (no restriction on provider type): Diabetic nephropathy, ESRD, Chronic renal failure (CRF), Chronic kidney disease (CKD), Renal insufficiency, Proteinuria, Albuminuria, Renal dysfunction, Acute renal failure (ARF), Dialysis, hemodialysis or peritoneal dialysis.

-Evidence of ACE inhibitor/ARB therapy. Documentation in the medical record must include evidence that the member received ACE inhibitor/ARB therapy during the measurement year. Any of the following meet criteria:, Documentation that a prescription for an ACE inhibitor/ARB was written during the measurement year, Documentation that a prescription for an ACE inhibitor/ARB was filled during the measurement year, Documentation that the member took an ACE inhibitor/ARB during the measurement year.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18-75 years of age by the end of the measurement year who had a diagnosis of diabetes (type 1 or type 2) during the measurement year or the year prior to the measurement year.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Patients with diabetes can be identified two ways:
-CLAIM/ENCOUNTER DATA: Patients who had two face-to-face encounters, in an inpatient setting or nonacute inpatient setting, or ED setting on different dates of service, with a diagnosis of diabetes, or one face-to-face encounter in an acute inpatient, with a diagnosis of diabetes, during the measurement year or the year prior to the measurement year. Organizations may count services that occur over both years.

*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

-PHARMACY DATA: Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year.

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES (TABLE CDC-A):
Alpha-glucosidase inhibitors:
Acarbose, Miglitol

Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metaformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metaformin-saxagliptin, Metformin-sitagliptin , Sitagliptin-simvastatin

Insulin:
Insulin aspart, Insulin aspart-insulin aspart protamine, insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitagliptin

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclusions (optional):
-Exclude patients who did not have a diagnosis of diabetes, in any setting, AND who had a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year
-Exclude patients 65 and older with an advanced illness condition and frailty

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.)*
ADMINISTRATIVE CLAIMS:
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

Exclude patients who do not have a diagnosis of diabetes (Diabetes Value Set), in any setting, during the measurement year or the year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes (Diabetes Exclusions Value Set), in any setting, during the measurement year or the year prior to the measurement year.

Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

MEDICAL RECORD:
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and had a diagnosis of polycystic ovaries any time in the patient's history through December 31 of the measurement year.
OR
-Exclusionary evidence in the medical record must include a note indicating the patient did not have a diagnosis of diabetes, in any setting, during the measurement year or the year prior to the measurement year and a diagnosis of gestational or steroid-induced diabetes, in any setting, during the measurement year or the year prior to the measurement year.

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

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 = Higher 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.)*
STEP 1. Determine the eligible population. To do so, identify patients who meet all the specified criteria.
-AGES: 18-75 years as of December 31 of the measurement year.
-EVENT/DIAGNOSIS: Identify patients with diabetes in two ways: by claim/encounter data and by pharmacy data.
Claim/Encounter Data:
-Patients who had at least two outpatient visits, observation visits, ED visits or nonacute inpatient encounters on different dates of service, with a diagnosis of diabetes. Visit type need not be the same for the two visits.
-Patients with at least one acute inpatient encounter with a diagnosis of diabetes.
*SEE ATTACHED EXCEL FILE FOR CODE VALUE SETS INCLUDED IN QUESTION S.2B

Pharmacy Data:
Patients who were dispensed insulin or hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year.
*SEE PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES IN QUESTION S.7

STEP 2. Determine the number of patients in the eligible population who had a recent nephropathy screening or monitoring test or evidence of nephropathy or treatment of nephropathy during the measurement year through the search of administrative data systems.
STEP 3. Identify patients with a nephropathy screening or monitoring test or evidence of nephropathy.
STEP 4. Identify the most recent nephropathy screening or monitoring test or evidence of nephropathy during the measurement year (numerator compliant). Identify the missing nephropathy screenings or monitoring tests or no evidence of nephropathy (not

numerator compliant).
STEP 5. Exclude from the eligible population patients from step 2 for whom administrative system data identified an exclusion to the service/procedure being measured.
*SEE DENOMINATOR EXCLUSION CRITERIA IN QUESTION S.8
STEP 6. Calculate the rate (number of patients with nephropathy screening or monitoring test or evidence of nephropathy during the measurement year or year prior?).

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data, Other, Paper Medical 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.
This measure uses a combination of administrative claims data and medical records. A nephropathy screening or monitoring test or evidence of nephropathy during the measurement year can be identified by the following administrative data:

-A nephropathy screening or monitoring test (Urine Protein Tests Value Set).
-Evidence of treatment for nephropathy or ACE/ARB therapy (Nephropathy Treatment Value Set).
-Evidence of stage 4 chronic kidney disease (CKD Stage 4 Value Set).
-Evidence of ESRD (ESRD Value Set).
-Evidence of kidney transplant (Kidney Transplant Value Set).
-A visit with a nephrologist, as identified by the organization's specialty provider codes (no restriction on the diagnosis or procedure code submitted).
-At least one ACE inhibitor or ARB dispensing event (ACE Inhibitor/ARB Medications List).

Medical record documentation includes:
-A urine test for albumin or protein. At a minimum, documentation must include a note indicating the date when a urine test was performed, and the result or finding. Any of the following meet the criteria: 24-hour urine for albumin or protein, timed urine for albumin or protein, spot urine (e.g., urine dipstick or test strip) for albumin or protein, urine for albumin/creatinine ratio, 24-hour urine for total protein, random urine for protein/creatinine ratio.
-Documentation of a visit to a nephrologist.
-Documentation of a renal transplant.
-Documentation of medical attention for any of the following (no restriction on provider type): diabetic nephropathy, ESRD, chronic renal failure (CRF), chronic kidney disease (CKD), renal insufficiency, proteinuria, albuminuria, renal dysfunction, acute renal failure (ARF), dialysis, hemodialysis or peritoneal dialysis.
-Evidence of ACE inhibitor/ARB therapy. Documentation in the medical record must include evidence that the member received ACE inhibitor/ARB therapy during the measurement year. Any of the following meet criteria: Documentation that a prescription for an ACE inhibitor/ARB was written during the measurement year, Documentation that a prescription for an ACE inhibitor/ARB was filled during the measurement year, Documentation that the member took an ACE inhibitor/ARB during the measurement year.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Health Plan

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

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

2. Validity – See attached Measure Testing Submission Form
NQF_Testing_0062_Nephropathy_7.1-636588879996129718.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0062
Measure Title: Comprehensive Diabetes Care: Medical Attention for Diabetic Nephropathy
Date of Submission: 3/5/2018
Type of Measure:

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

Instructions

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

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ 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? 2010-2012

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*)
Health Plan Level

We calculated the measure score reliability and construct validity from HEDIS data that included 416 commercial health plans, 500 Medicare health plans, and 194 Medicaid health plans. The sample included all commercial, Medicare, and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Physician Level

We also calculated measure score reliability from physician/practice level data from the NCQA Diabetes Recognition Program (DRP) that included 3676 physicians. Construct validity was calculated with data from a sample of 653 physicians/practices.

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*)

2012 data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicaid, Medicare). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible plans for the measure across health plans.

HEDIS Health Plan

Product Type	Number of Plans	Median Number of Eligible Patients per Plan
Commercial HMO	218	2,804
Commercial PPO	198	6,445
Medicaid HMO	194	1,846
Medicare HMO	349	1,586
Medicare PPO	151	1,527

NCQA's Diabetes Recognition Program currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes. Below is a description of the sample. It includes the number of physicians and practices reporting on this measure in the DRP program in 2012.

Physician Level

Analysis	Number of physicians	Median Denominator Size
Reliability	3,676	25
Construct Validity	653	25

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

Reliability:

Reliability of the health plan measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Reliability of the physician/practice level measure in the DRP was tested using a beta-binomial calculation. This analysis included the entire DRP sample (described above).

Validity:
Validity of the health plan measure was demonstrated through construct validity using the entire HEDIS data sample (described above) and through a systematic assessment of face validity with expert panels.

Validity was demonstrated through construct validity using data from a sample of 653 physicians/practices and through a systematic assessment of face validity with expert panels.

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

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*)
Reliability Testing of Performance Measure Score:
Reliability was estimated by using the beta-binomial model for the health plan measure and physician/practice level DRP measure. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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*)

Health Plan Level

Product Type	Reliability per Beta Binomial Model
Commercial	1.00
Medicare	0.97
Medicaid	0.97

Physician Level

Product Type	Reliability per Beta Binomial Model
Diabetes Recognition Program	0.90

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?*)

Health Plan Level

The values for the beta-binomial statistic across all product lines for the health plan level measure suggest the measure has high reliability.

Physician Level

The value for the beta-binomial statistic for the physician level measure suggest the measure has high reliability.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Method of Testing Construct Validity – Health Plan Level

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality hypothesized to be related, which are listed below.

- HbA1c Testing
- Hemoglobin (HbA1c) Poor Control (>9%)
- Eye Examination (Eye Exam)

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second

variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of Testing Construct Validity – Physician Level

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality in NCQA's Diabetes Recognition Program hypothesized to be related, which are listed below.

- Eye Exam
- Smoking and Tobacco Use and Cessation and Treatment Assistance (Smoking Cessation)
- Foot Examination (Foot Exam)

We tested the correlations using the Pearson correlation test described above.

Method of Assessing Face Validity – Health Plan Level

We describe below NCQA's process for both measure development and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA's Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assure that the measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.
STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed periodically, based on changes in evidence and guidelines. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support (PCS) portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures. Over the past four years, NCQA has received and responded to an average of 39 inquiries per year on this measure.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2 and in other relevant NCQA programs.

Method of Assessing Face Validity - Physician Level
The physician level measure was tested for face validity with four panels of experts. The Diabetes Recognition Program (DRP) Advisory Committee included 7 experts in diabetes care including representation by clinicians, health plans, integrated health systems and research organizations; DMAP, CPM and the Clinical Programs Committee (CPC). NCQA's CPC's oversees the evolution of NCQA's recognition programs and related measures including the Diabetes Recognition Program, the Heart/Stroke Recognition Program, the Patient Centered Medical Home and Patient-Centered Specialty Practice Recognition Program, among others. The CPC includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 18 members. The CPC is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of clinical recognition programs. CPC members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
The results from construct validity testing of the health plan level measure are presented by product line in Tables 1a, 1b, and 1c below.

Table 1a. Correlations among Diabetes Measures in Commercial Health Plans - 2012

	Pearson Correlation Coefficients		
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Eye Exam
CDC – Medical Attention for Diabetic Nephropathy	0.76	-0.61	0.72

Note: All correlations are significant at $p<0.0001$

Table 1b. Correlations among Diabetes Measures in Medicaid Health Plans - 2012

	Pearson Correlation Coefficients		
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Eye Exam
CDC – Medical Attention for Diabetic Nephropathy	0.56	-0.52	0.45

Note: All correlations are significant at $p<0.0001$

Table 1c. Correlations among Diabetes Measures in Medicare Health Plans - 2012

	Pearson Correlation Coefficients		
	HbA1c Testing	HbA1c Poor Control (>9.0%)	Eye Exam
CDC – Medical Attention for Diabetic Nephropathy	0.42	-0.29	0.38

Note: All correlations are significant at $p<0.0001$

Construct Validity – Physician Level
Table 2a below provides the results from construct validity testing of the physician level measure.

Table 2a. Correlations among HbA1c Measures in the NCQA Diabetes Recognition Program - 2012

	Pearson Correlation Coefficients		
	Eye Exam	Smoking Cessation	Foot Exam
CDC – Medical Attention for Diabetic Nephropathy	0.26	0.55	0.29

Note: All correlations are significant at $p<0.0001$

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?*)
Construct Validity – Health Plan Level
Across all product lines, the correlations are moderate to strong and statistically significant. These results confirmed the hypothesis that the diabetes measures are correlated with each other. Coefficients with absolute value of less than .3 are generally considered indicative of weak associations. Absolute values of .3 to .59 are considered moderate associations, absolute values of .6 to .69 indicate a strong positive relationship, and absolute values of .7 or higher indicate a very strong positive relationship. These correlation results suggest that at the plan level the measure has sufficient validity.

Note: Correlation values with the HbA1c Poor Control measure are all negative because it is a "lower is better quality" measure, while the other measures are all "higher is better".

Construct Validity - Physician Level
At the physician level, the *CDC – Medical Attention for Diabetic Nephropathy* measure has a moderate correlation with the *Smoking and Tobacco Use and Cessation and Treatment Assistance* measure in the Diabetes Recognition Program. The correlation between the *Eye Exam* and *Foot Exam* measures is lower and indicates a slightly weaker association. Overall these correlation results suggest that the physician level measure has sufficient validity.

Face Validity – Health Plan Level
NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that the *CDC – Medical Attention for Diabetic Nephropathy* measure is measuring what it intends to measure. The results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

Face Validity – Physician Level
These results indicate that the multiple experts, stakeholders and NCQA's Clinical Programs Committee concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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*)
Testing was not performed for the excluded sample.

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*)
Testing was not performed for the excluded sample.

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*)
Testing was not performed for the excluded sample.

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

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

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**
☐ **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*)
N/A

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample, size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans performance is significantly different from each other.

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*)

Health Plan Level - 2012

Product Type	N	Mean (%)	St Dev (%)	P10th (%)	P25th (%)	P50th (%)	P75th (%)	P90th (%)	IQR (%)	P value
Commercial HMO	218	84.25	5.66	77.78	80.50	84.40	88.18	90.79	7.68	<0.05
Commercial PPO	198	78.59	6.53	70.26	75.69	79.34	82.73	85.59	7.04	<0.05
Medicaid HMO	194	78.41	7.31	69.76	75.00	79.28	82.74	85.85	7.74	<0.05
Medicare HMO	349	89.96	5.15	85.16	87.83	90.28	92.70	95.07	4.87	<0.05
Medicare PPO	151	88.30	3.66	84.67	86.37	88.32	90.51	92.19	4.41	<0.05

N = total number of plans reporting data
IQR: Interquartile range
p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

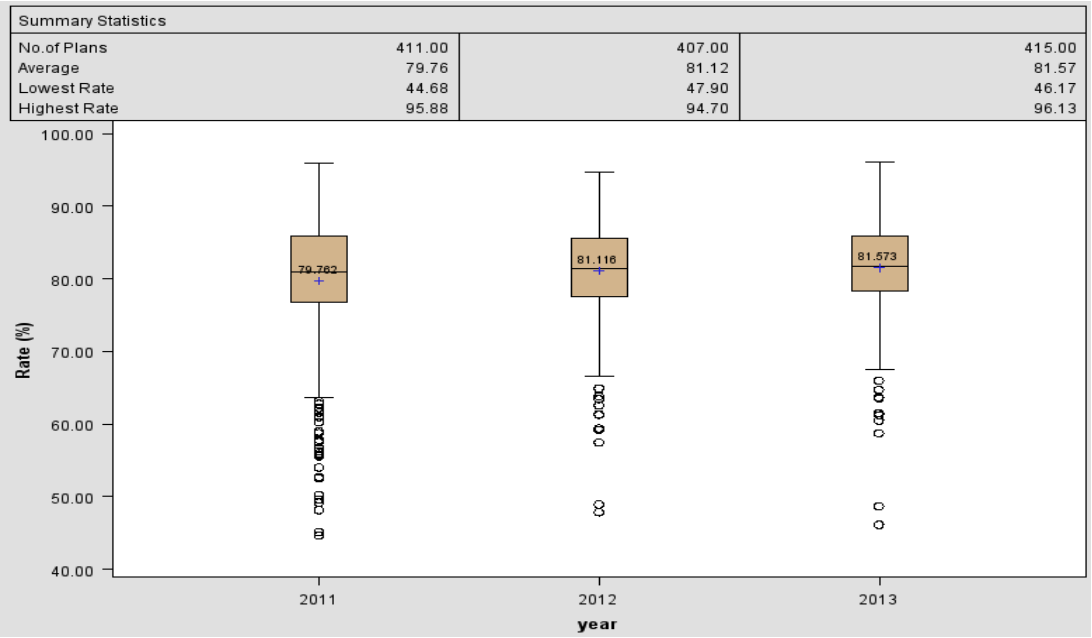
Physician Level - 2012

N (# of clinicians)	Mean (%)	St Dev (%)	P10th (%)	P25th (%)	P50th (%)	P75th (%)	P90th (%)	IQR (%)	P value
3676	86.48	17.73	74.00	84.00	92.00	96.00	100.00	12.00	<0.05

IQR: Interquartile range
p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

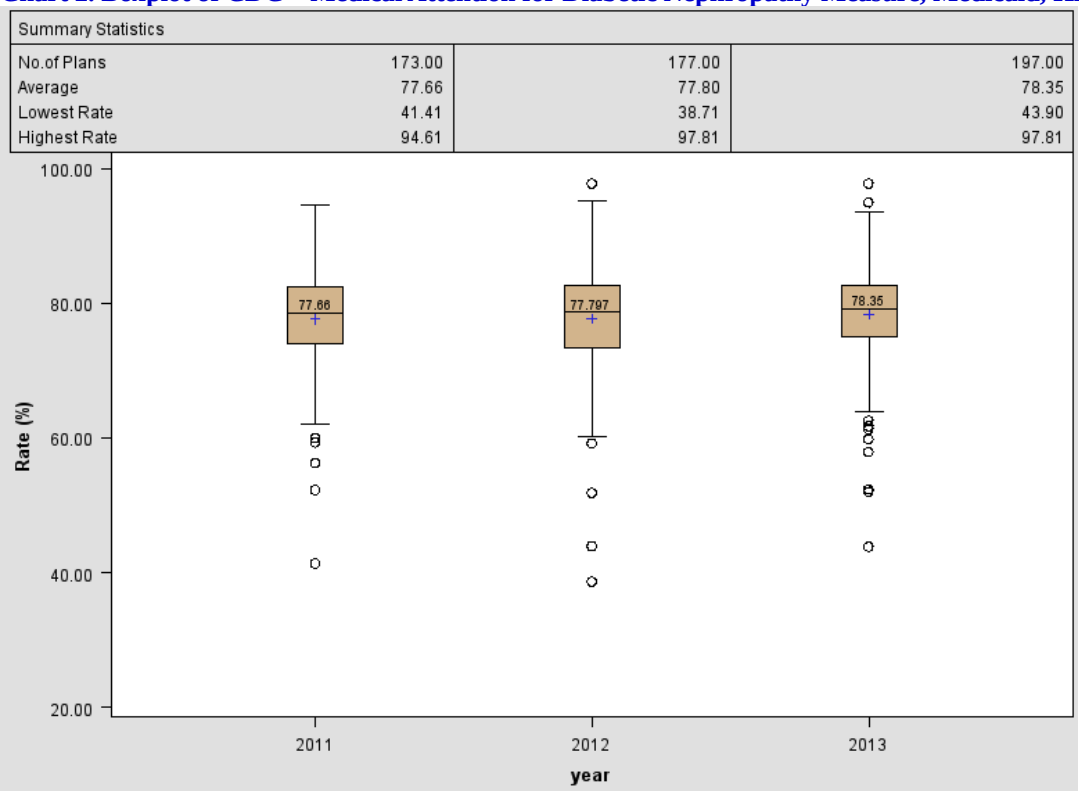
Health Plan

Chart 1. Boxplot of CDC – Medical Attention for Diabetic Nephropathy Measure, Commercial, HEDIS 2011-2013*



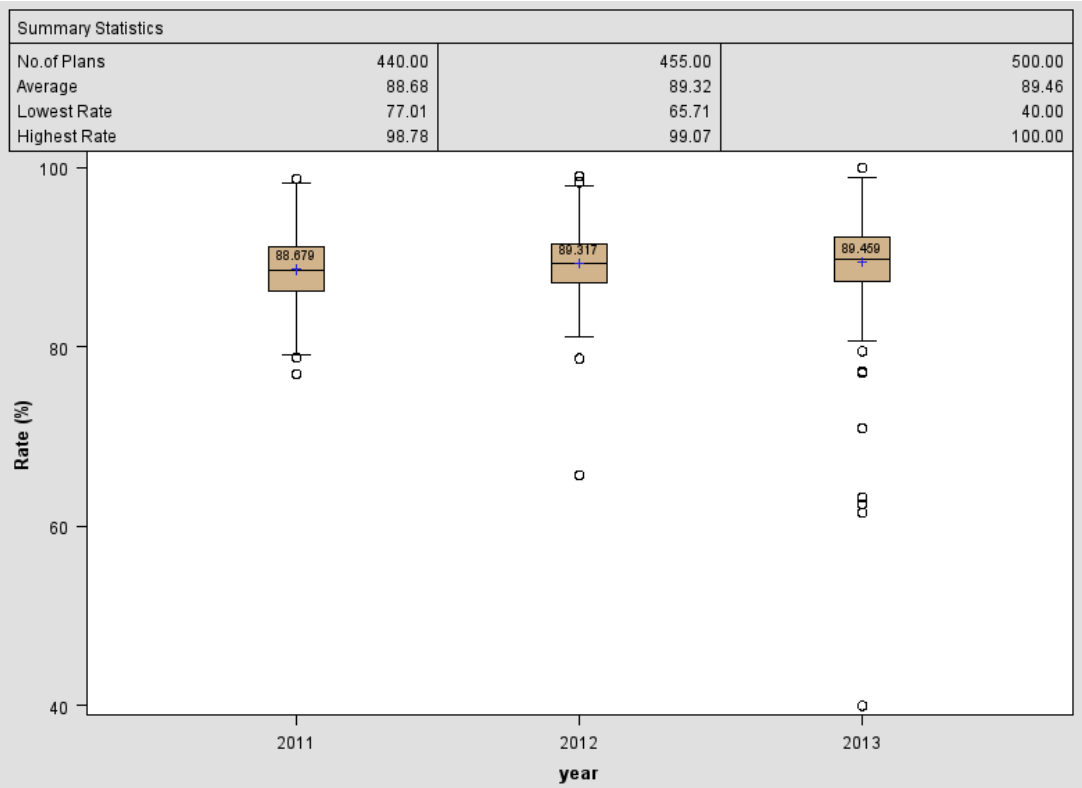
* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Chart 2. Boxplot of CDC – Medical Attention for Diabetic Nephropathy Measure, Medicaid, HEDIS 2011-2013*



* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

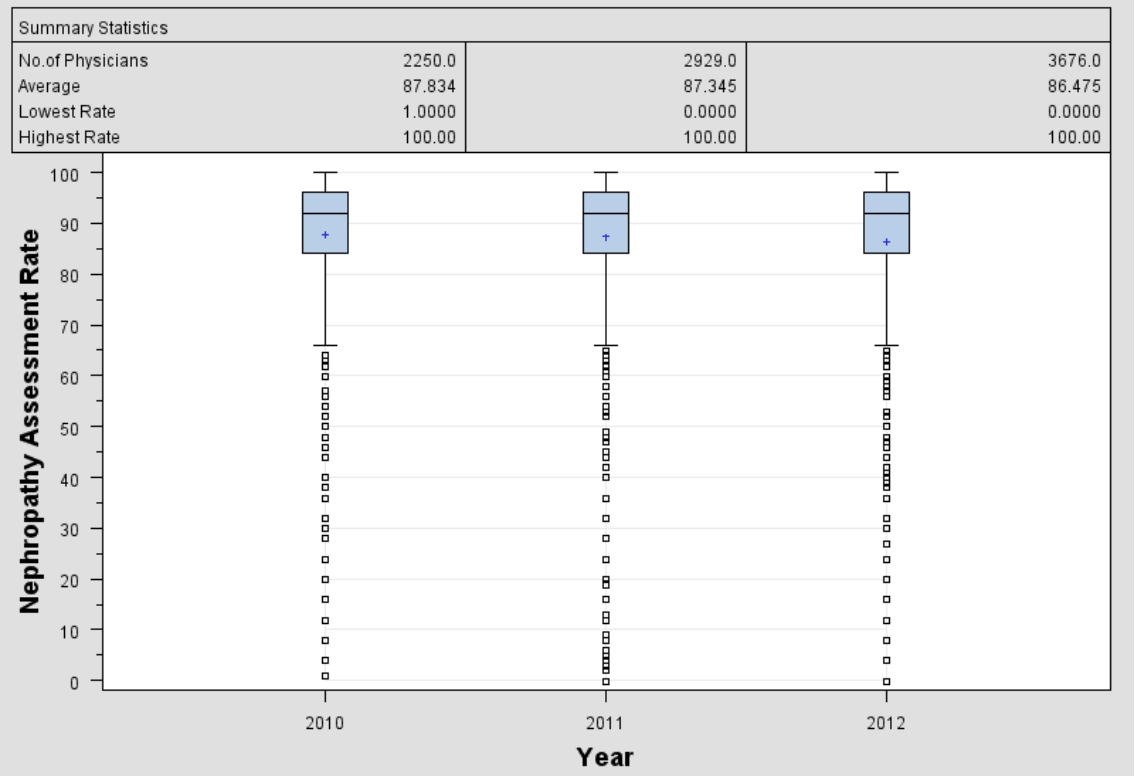
Chart 3. Boxplot of CDC – Medical Attention for Diabetic Nephropathy Measure, Medicare, HEDIS 2011-2013*



* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Physician Level

Chart 4. Boxplot CDC – Medical Attention for Diabetic Nephropathy Measure, Diabetes Recognition Program, 2010-2012



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?*)

Health Plan Level

Across all product lines, the difference between the 25th (better performance) and 75th percentile is statistically significant. Overall, these results suggest there are meaningful differences in performance.

Physician Level

The difference between the 25th and 75th percentile is statistically significant, suggesting there are meaningful differences in performance.

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

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 non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
This measure is collected with a complete sample.

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*)
This measure is collected with a complete sample.

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

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), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
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**.

Some data elements are in defined fields in electronic sources

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).

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, and paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease

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

Attachment:

3c. Data Collection Strategy

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

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

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

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
N/A

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 Health Plan Ranking http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Physician Quality Reporting System (PQRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ Health Plan Ranking http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Physician Quality Reporting System (PQRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ Payment Program IHA California Pay for Performance http://www.iha.org/manuals_operations_2014.html IHA California Pay for Performance http://www.iha.org/manuals_operations_2014.html Regulatory and Accreditation Programs NCQA Accreditation; Accountable Care Organizations (ACO) http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx NCQA Accreditation; Accountable Care Organizations (ACO) http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx

	Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

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

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

HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population

CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

INTEGRATED HEALTHCARE ASSOCIATION (IHA) CALIFORNIA PAY FOR PERFORMANCE: This measure is used in the California P4P program which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of eight health plans representing 10 million insured persons. IHA is responsible for collecting data, deploying a common measure set, and reporting results for approximately 35,000 physicians in nearly 200 physician groups. This program represents the longest running U.S. example of data aggregation and standardized results reporting across diverse regions and multiple health plans. California consumers benefit from the availability of standardized performance results from a common measure set, which are available to the public through the State of California, Office of the Patient Advocate

ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

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?*)
N/A

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.*)
N/A

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Questions received through the Policy Clarification Support system have generally centered around clarification on types of lab tests that are considered screening or monitoring for nephropathy such as creatinine/glomerular filtration rate and urinalysis or documentation of history of mico albuminuria or if patients must be on an ACE/ARB the entire measurement year to be counted in the measure.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the PQRS and the Health Plan Rankings program.

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.**
We have provided minor clarifications about the measure during the annual update process in order to address questions received through the Policy Clarification Support system.

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.
From 2014 to 2016, performance rates for this measure have increased for all product lines (Commercial, Medicaid, and Medicare). Of the plans, the highest performance continues to be seen in the Medicare population. In 2016, Medicare plans had a performance rate of 97 percent while Commercial and Medicaid has around 90 percent (see section 1b.2 for summary of data from commercial, Medicaid, and Medicare Health Plans). These data are nationally representative.

Since 2015, there has been a decrease in the number of reporting physicians seeking recognition in NCQA's DRP program (see summary data in 1b.2). However, we are pleased that rates in performance have gone up slightly.

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.
There were no identified unexpected findings during testing or since implementation of this measure

4b2.2. Please explain any unexpected benefits from implementation of this measure.
There were no identified unexpected findings during testing or since implementation of this measure

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.
No

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?
No

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

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

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

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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.
DIABETES EXPERT PANEL:
Bill Herman (Chair), MD, Univ. of Michigan Health System
David Aron, MD, Department of Veteran's Affairs
James Fain, PhD, RN, University of Massachussetts
Jerry Cavallerano, OD, Beetham Eye Institute
John Thompson, MD, Retina Specialists
Judith Fradkin, MD, NIDDK/NIH
Lynne Levitsky, MD, Massachusetts General Hospital
Mark Cziraky, PharmD, Healthcore
Richard Hellman, MD, Private Practice, Diabetes & Endocrinology
Seth Rubenstein, DPM, Reston Hospital Center, INOVA Fair Oaks Hospital
Stephen Fadem, MD, Baylor College of Medicine
Ted Ganiats, MD, Univ. of California, San Diego
Nancy Van Vessem, MD, Capital Health Plan

HEDIS EXPERT CODING PANEL
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Aqurate Health Data Management, Inc.
Nelly Leon-Chisen, RHIA, American Hospital Association
Alec McLure, MPH, RHIA, CCS-P, Verscend Technologies
Michele Mouradian, RN, BSN, Change HealthCare
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, WellCare Health Plans, Inc.

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management

Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

CLINICAL PROGRAMS COMMITTEE
Randall Curnow, MD, MBA, FACP, FACHE, FACPE (Chair), TriHealth
Suzanne Berman, MD, FAAP, Plateau Pediatrics
Brooks Daveman, MPP, Tennessee Division of Health Care Finance and Administration
Marcus Friedrich, MD, MBA, FACP, New York State Department Health Empire State Plaza, Coming Towne
Jennifer Gutzmore, MD, Cigna
Melissa Hogan, MPH, Aon
Adriana Matiz, MD, FAAP, Ambulatory Care Network
Lisa Morrise, Marts, LAM Professional Services, LLC
Deborah Murph, MBA, BSN, RN, Cherokee Health Systems
Amy Nguyen Howell, MD, MBA, CAPG
Marc Rivo, MD, Population Health Innovations
Julie Schilz, BSN, MBA, Anthem
Pamela Slaven-Lee, DNP, FNP-C, CHSE, The George Washington University School of Nnursing
Lina Walker, PhD, AARP

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 1999

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.6 **Copyright statement:** The performance measures and specifications were developed by and are owned by the National Committee for Quality Assurance ("NCQA"). The performance measures and specifications are not clinical guidelines and do not establish a standard of medical care. 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. NCQA holds a copyright in these materials and can rescind or alter these materials at any time. These materials may not be modified by anyone other than NCQA. Anyone desiring to use or reproduce the materials without modification for an internal, quality improvement non-commercial purpose may do so without obtaining any approval from NCQA. All other uses, including a commercial use and/or external reproduction, distribution and publication must be approved by NCQA and are subject to a license at the discretion of NCQA.

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

Content reproduced with permission from HEDIS, Volume 2: Technical Specifications for Health Plans. To purchase copies of this publication, including the full measures and specifications, contact NCQA Customer Support at 888-275-7585 or visit www.ncqa.org/publications.

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care