NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 0086

Corresponding Measures: 0086e

De.2. Measure Title: Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation

Co.1.1. Measure Steward: PCPI Foundation

De.3. Brief Description of Measure: Percentage of patients aged 18 years and older with a diagnosis of primary open-angle glaucoma (POAG) who have an optic nerve head evaluation during one or more office visits within 12 months

1b.1. Developer Rationale: Glaucoma is a group of diseases that damage the eye's optic nerve and can result in vision loss and blindness. In 2011, 2.71 million persons in the U.S. had primary open-angle glaucoma (POAG) and in 2050, an estimated 7.32 million persons will have POAG (1). Furthermore, a 2006 study estimated that the total financial burden of major visual disorders among U.S. residents aged 40 years or older was $35.4 billion in 2004: $16.2 billion in direct medical costs, $11.1 billion in other direct costs, and $8 billion in productivity losses. Of the direct medical costs, approximately $2.9 billion was attributable to glaucoma (2). It is imperative that evidence-based care be delivered to all glaucoma patients.

According to the most recent guidelines, changes in the optic nerve are one of two characteristics which currently define progression and thus worsening of glaucoma disease status (the other characteristic is visual field). Examination of the optic nerve head (ONH) and retinal nerve fiber layer (RNFL) provides valuable structural information about glaucomatous optic nerve damage. Visible structural alterations of the ONH or RNFL and development of parapapillary choroidal atrophy in early glaucoma may precede the onset of visual field defects. Careful study of the optic disc neural rim for small hemorrhages is important because these hemorrhages sometimes herald focal disc damage and visual field loss, and they may signify ongoing optic nerve damage in patients with glaucoma (3). Despite evidence emphasizing the value of an optic nerve evaluation, there is a gap in documentation patterns of the optic nerve for both initial and follow-up care.

This measure is intended to promote examination and documentation of the structure and function of the optic nerve, and to monitor and detect disease progression among POAG patients. This measure should lead to the desired health outcome of preservation of one's visual function and ultimately the maintenance of quality of life for the patient.

1. Vajaranant, T. S., Wu, S., Torres, M., & Varma, R. (2012). The Changing Face of Primary Open-Angle Glaucoma in the United States: Demographic and Geographic Changes From 2011 to 2050. American Journal of Ophthalmology, 154(2). doi:10.1016/j.ajo.2012.02.024

2. Rein, D. B., Zhang, P., & Wirth, K. (2006). The Economic Burden of Major Adult Visual Disorders in the United States. Archives of Ophthalmology, 124(12), 1754-1760. doi:10.1001/archopht.124.12.1754

3. Prum, B. E., Rosenberg, L. F., Gedde, S. J., Mansberger, S. L., Stein, J. D., Moroi, S. E., . . . Williams, R. D. (2015). Primary Open-Angle Glaucoma Preferred Practice Pattern® Guidelines. Ophthalmology, 123(1). doi:10.1016/j.ophtha.2015.10.053

S.4. Numerator Statement: Patients who have an optic nerve head evaluation during one or more office visits within 12 months

S.6. Denominator Statement: All patients aged 18 years and older with a diagnosis of primary open-angle glaucoma

S.8. Denominator Exclusions: Denominator Exceptions:

Documentation of medical reason(s) for not performing an optic nerve head evaluation

De.1. Measure Type: Process

S.17. Data Source: Claims, Registry Data

S.20. Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: May 01, 2007 **Most Recent Endorsement Date:** Nov 04, 2015

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?

Preliminary Analysis: Maintenance of Endorsement

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 2015:

- Brief background: This measure looks at patients 18+ years of age with glaucoma who had optic nerve head evaluation within a 12 month period. The measure was reviewed in 2015 by the Eye Care and Ear, Nose and Throat Conditions Project, Eyes and Glaucoma disease Subtopic.
- Developer provided an updated logic model depicting the relationship between optic nerve evaluation, adjustments in therapy and enhanced patient outcomes.
- The developer associated evaluation of optic nerve structure and function with improvements in Primary Open Angle Glaucoma management and improvements in quality of life, stating that optic nerve head evaluation provides information that informs therapeutic goals to preserve visual function The developer reports that "People with vision loss are more likely to report depression, diabetes, hearing impairment, stroke, falls, cognitive decline, and premature death. Decreased ability to see often leads to the inability to drive, read, keep accounts, and travel in unfamiliar places, thus substantially compromising quality of life."
- The developer provided updated clinical practice guidelines from the AAO 2015 Preferred Practice Pattern Guidelines which states eye assessment for glaucoma should include "Optic nerve head and retinal nerve fiber layer examination". The AAO classifies the evidence as both Level III- moderate quality (on GRADE) and I+ on SIGN (including well-conducted meta-analyses, systematic reviews of RTCs, or RCTs with a low risk of bias). The care process is level A:strong recommendation or most important.
- Since the guidelines did not include QQC of the evidence, the developer analyzed studies supporting the current guidelines. The developer found 38 studies supporting current recommended guidelines by AAO; 32 of the 38 studies were observational studies (including 7 descriptive studies, 23 cohort studies, and 2 case/control studies), and 6 were RCT including the Early Manifest Glaucoma Trial and the European Glaucoma Prevention Study. The studies corroborated on the importance of evaluating the optic head nerve as an indicator for evaluationg glaucoma and glaucoma damage.

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 noted that there have been no changes in evidence; however, they have updated the form to capture the current language in the most recent guideline. Optic nerve head assessment remains one of two exams used in evaluating the status of glaucoma.

Questions for the Committee:

o The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review, and has made limited adjustments to the evidence submitted. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?

Guidance from the Evidence Algorithm

Algorithm 1. – Process measure (box 3) → QQC provided (Box 4) → Box 5 Used to determine evidence rating.

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.

Data from CMS QPP PQRS and AOA:

Year	Modality	Average Performance Rate
2017	QPP	90.17%
2016	PQRS	91.6%
2015	PQRS	91%
2014	PQRS	94%
2013	PQRS	95.4%

Year / Modality / Reporting Rate (percentage of those eligible to report on measure)

Year	Modality	Average Performance Rate
2017	QPP	85.06%
2016	PQRS	36.9%
2015	PQRS	43.7%
2014	PQRS	44.5%
2013	PQRS	38.1%

American Optometric Association (AOA) Measures and Outcomes Registry for Eyecare (MORE) Registry/QCDR, for this measure:

Year/Modality/Average Performance Rate

Year	Modality	Average Performance Rate
2018	QCDR	75%
2017	QCDR	53%

Disparities

- Althought this measure is used in Federal reporting programs, the developer noted that disparities data are not available for analysis and report.
- The developer reported African Americans age 40 and older are at highest risk of developing glaucoma and make up 21% of the Medicaid population. However, a longitudinal cohort study documented disproportionately fewer number of Medicaid users versus commercial health insurance users receiving glaucoma testing, indicating racial disparities amongst those who receive glaucoma testing following initial diagnosis. According to the developer, Medicaid recipients were 234% more likely to not receive any glaucoma testing in the 15 months following initial diagnosis.
- A retrospective cohort study found that women were 24% less likely to undergo treatment than men.

Questions for the Committee:

- Performance on this measure over time appears to have consistently gone down. Does this represent a concern for the Committee for the importance of this measure?
- Does the developer's evidence suggest an opportunity to address healthcare disparities using this measure?

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

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

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

- This is a process measure. The evidence applies directly to the process of glaucoma management. Evaluation of the optic nerve structure is one of the two methods to monitor glaucoma progression (along with visual field assessment). Periodic evaluation of optic nerve structure and visual field allow providers to make wise treatment decisions. No new studies change the evidence base for the measure as written.
- The evidence applies directly and there is no significant new evidence for maintenance.
- It applies directly to the disease state of glaucoma. The structure and process of an ophthalmic exam in patient with POAG relates to longitidual optic nerve head evaluation. This should be done annually and not necessary at every visit.
- The measure continues to address best practice based on current evidence. The evidence noted directly ties to the process function of the measure.

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

- Performance data suggests 90% performance in 2017 (95% in 2013) and a jump in reporting rate from ~40% to 85% in 2017. An optometry performance registry showed improvement in average performance rate from 53% to 75% from 2017-2018. A gap in care likely exists. Disparity data was not provided directly, but it can be inferred that improvement in performance will address underlying disparity that exists in glaucoma
- Although the performance gap has been traditionally high, the percentage has trended lower suggesting a meaningful gap in performance. Disparity data was not reported although some evidence suggests a disparity with respect to race and sex.
- The results show likely a non-significant change in outcome patterns from when the measure was initiated over the past few years. With regards to the medicaid patients not receiving adequate care that might be related to their overall compliance with exams. I don't believe that anything could be added from an inclusion/exclusion to specify for this. No specific data on population subgroups was provided from the reporting data.
- Yes. The performance ratings for this measure continues to decline (decrease from 95.4% in 2013 to 90.17% in 2017). The developer was able to show disparities from a studies perspective which displays a gap when comparing Medicaid members to commercial population and across ethnicities. However, it was not captured on the data displayed from the federal reporting program which displays their ratings.

Criteria 2: Scientific Acceptability of Measure Properties

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

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

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

Evaluators: Primary Care and Chronic Illness project team staff

Link A (Project Team staff)

Reliability:

- The developer conducted updated performance measure score reliability testing using data at the claims and registry level of analysis listed below. Reliability testing was performed by using a beta-binomial model (i.e. signal to noise).
- Measure was testing using pooled data for individual and group clinicians, but is specified for both. NQF requires separate analysis for these two types of providers. Therefore, this measure is considered insufficient; testing must be performed to specification per NQF criteria.
- Developer has specified the measure for outpatient, post acute care and domicillary settings, but these analyses were not conducted separately, nor is the care setting clearly articulated in the submission.
- Claims Data (PQRS): The data is from 13,685 providers reporting on this measure through the claims reporting option for CMS's PQRS. From January 2016 through December 2016. This dataset reflects a combination of individual provider data and group data and the analysis of the data as a whole is reflected throughout this submission. Of those, 13,377 providers had all the required data elements and at least one quality reporting event for a total of 499,801 quality events. The remainder of providers had 0 quality reporting events after accounting for exceptions. For this measure, 98.0 percent of providers are included in the analysis, and the average number of quality reporting events is 37. The range of quality reporting events for 13,377 providers included is from 1 to 1,368.
- Note that these providers were not separated according to individual and group levels of analysis, although the measure is specified this way.
- Registry Data (PQRS): The data is from 1,591 providers reporting on this measure through the registry reporting option for CMS's PQRS. From January 2016 through December 2016. This dataset reflects a combination of individual provider data and group data and the analysis of the data as a whole is reflected throughout this submission. 1,591 providers had all the required data elements and at least one quality reporting event for a total of 152,753 quality events. For this measure, 100 percent of providers are included in the analysis, and the average number of quality reporting events is 96. The range of quality reporting events for 1,591 providers included is from 1 to 2,712.

The developer had the following results for their two data samples.

- Claims Data (PQRS): The average reliability including providers with at least one quality reporting event is 0.96. The developer also evaluated reliability at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} cut points, which ranged from 0.95 at 10^{th} percentile and 1.0 at the remaining cut points.

- Registry Data (PQRS): The average reliability including providers with at least one quality reporting event is 0.95. The developer also evaluated reliability at the 10th, 25th, 50th, 75th, and 90th cut points, which ranged from 0.86 at 10th percentile, 0.98 at 25th percentile, and 1.0 at the remaining cut points.

For signal to noise, a reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. Both the reliability results for claim and registry data had very high reliability.

Validity:

Empirical Validity:

- The developer did convergent validity testing with Pearson's correlation coefficients and compared the performance of NQF 0086 with PQRS#117 Diabetes: Eye Exam. This is an appropriate method for empirical validity of the measure.
- The results of empirical validity was a moderate positive correlation at the registry level (0.57). However, result was a weak positive correlation at the claim level (0.22). The developer noted PQRS 117 is only existing measure they could correlate with 0086.

Face Validity:

- The developer previously did face validity testing in 2013. The developer did face validity of the measure score with the expert panel for the measure. The expert panel included 16 members.
- For face validity testing results from 2013, the mean rating was 4.56 and 87.50% of TEP either agreed or strongly agreed that this measure can accuarately distinguid good and poor quality.

Questions for the Committee regarding reliability:

- Note: This measure is specified for both group and individual clinicians, which by NQF criteria must be tested separately. The testing was performed with these two groups pooled together, and hence is not tested to specifications.
- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Committee should discuss the reliability scoring of the measure.

Questions for the Committee regarding validity:

- Note: This measure is specified for both group and individual clinicians, which by NQF criteria must be tested separately. The testing was performed with these two groups pooled together, and hence is not tested to specifications.
- In correlation testing, the registry measure scored moderate while the claims measure scored low. Would the Committee prefer to vote on these data sources separately for endorsement?
- The measure does not include a list of exclusions, instead relying on clinicians to appropriately exclude patients that should not recieive this exam for medical reasons. Does the Committee have any concerns about this or do you agree with their rationale?
- The developer states there was no missing data in their dataset, so they did not test for missing data. Is this a concern?
- Developer demonstrated moderate correlation with an external measure for registry data, but low correlation using claims. Committee to discuss these findings.
- Do you have any other concerns regarding the validity of the measure?

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

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

RATIONALE

- This measure is specified for both group and individual clinicians, which by NQF criteria must be tested separately.
- The testing was performed with these two groups pooled together, and hence is not tested to specifications.
- Developer demonstrated moderate correlation with an external measure for registry data, but low correlation using claims. Were the testing framework to be considered sufficient, this is still a low performing measure for validity for the claims data.

Evaluation A: Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0086

Measure Title: Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation

Type of measure:

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

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

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**

☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☒ **Registry Data**

☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☒ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**

☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

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

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

No concerns

RELIABILITY: TESTING

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

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

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

   Submission document: Testing attachment, section 2a2.2

   - This measure is specified for both group and individual clinicians, which by NQF criteria must be tested separately. The testing was performed with these two groups pooled together, and hence is not tested to specifications.
   - The developer conducted updated performance measure score reliability testing using data at the claims and registry level of analysis listed below. Reliability testing was performed by using a beta-binomial model (i.e. signal to noise).
   - Claims Data (PQRS): The data is from 13,685 providers reporting on this measure through the claims reporting option for CMS's PQRS. From January 2016 through December 2016. This dataset reflects a combination of individual provider data and group data and our analysis of the data as a whole is reflected throughout this submission. Of those, 13,377 providers had all the required data elements and at least one quality reporting event for a total of 499,801 quality events. The remainder of providers had 0 quality reporting events after accounting for exceptions. For this measure, 98.0 percent of providers are included in the analysis, and the average number of quality reporting events is 37. The range of quality reporting events for 13,377 providers included is from 1 to 1,368.
   - Registry Data (PQRS): The data is from 1,591 providers reporting on this measure through the registry reporting option for CMS's PQRS. From January 2016 through December 2016. This dataset reflects a combination of individual provider data and group data and the analysis of the data as a whole is reflected throughout this submission. 1,591 providers had all the required data elements and at least one quality reporting event for a total of 152,753 quality events. For this measure, 100 percent of providers are included in the analysis, and the average number of quality reporting events is 96. The range of quality reporting events for 1,591 providers included is from 1 to 2,712.
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - The developer had the following results for their two data samples.
     - Claims Data (PQRS): The average reliability including providers with at least one quality reporting event is 0.96. The developer also evaluated reliability at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} cut points, which ranged from 0.95 at 10^{th} percentile and 1.0 at the remaining cut points.
     - Registry Data (PQRS): The average reliability including providers with at least one quality reporting event is 0.95. The developer also evaluated reliability at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} cut points, which ranged from 0.86 at 10^{th} percentile, 0.98 at 25^{th} percentile, and 1.0 at the remaining cut points.
   - For signal to noise, a reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. Both the reliability results for claim and registry data had very high reliability.

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

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

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

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

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

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

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

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

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

    The measure is not tested to specifications, and thus is marked as insufficient.

    The reliability results otherwise indicated high reliability. However, would like to see results in future submissions; when data available, would like to see data from the MIPS program as PQRS is no longer in existence.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

- There is one denominator exception for medical reason for not performing an optic nerve head evaluation. However, the developer does not provide a list of reasons, but instead “relies on clinicians to link the exception with a specific reason for the decision not to perform the optic nerve evaluation required by the measure.” The developer notes that while this could raise concerns of inappropriate exclusions, “Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by providers, and it ameliorates perverse incentives to refuse care to "difficult" patients.”
- The developer did a denominator exception analysis for frequency across providers that suggested that exceptions do not represent a large proportion of the data.
  - Claims Data (PQRS): Amongst the 13,685 providers, there were a total of 10,911 exceptions reported. The average number of exceptions per provider in this sample is 1.0. The proportion of exceptions to patients is 0.02.

- Registry Data (PQRS): Amongst the 1,591 providers, there were a total of 2,036 exceptions reported. The average number of exceptions per provider in this sample is 1.0. The proportion of exceptions to patients is 0.01.

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

    Submission document: Testing attachment, section 2b4.

- The mean performance rates for this measure is extremely high at 0.97 Claims Data (PQRS) and 0.94 for Registry Data (PQRS).
  - Claims Data (PQRS) - The mean performance rate is 0.97 the median performance rate is 1.0 and the mode is 1.0. The standard deviation is 0.11. The range of the performance rate is 0.99, with a minimum rate of 0.01 and a maximum rate of 1.0. The interquartile range is 0.00 (1.00–1.00).
  - Registry Data (PQRS) – The mean performance rate is 0.94.The standard deviation is 0.15. The range of the performance rate is 0.95, with a minimum rate of 0.04 and a maximum rate of 1.0. The interquartile range is 0.03 (0.97–1.00).

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

    Submission document: Testing attachment, section 2b6.

- The developer notes there was no missing data in their dataset so they did not test for it. They note "missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance". However, they say there is no indication that any missing data might be systematic.

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

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

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

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

16d.**Risk adjustment summary:** N/A

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

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

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

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

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

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

N/A

VALIDITY: TESTING

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

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

    ☒ **Face validity**

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

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

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

    Submission document: Testing attachment, section 2b2.2

    Empirical Validity (new testing):

    The developer did convergent validity testing with Pearson's correlation coefficients and compared the performance of NQF 0086 with PQRS#117 Diabetes: Eye Exam . This is an appropriate method for empirical validity of the measure.

Correlation	Interpretation
0.80 – 1.00	Very Strong
0.60 – 0.79	Strong
0.40 – 0.59	Moderate
0.20 - 0.39	Weak
0 – 0.19	Very Weak

    Face Validity (2013):

    The developer previously did face validity testing in 2013. The developer did face validity of the measure score with the expert panel for the measure. The expert panel included 16 members.They were asked to rate their agreement with the following statement:

- "The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

  Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

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

    Submission document: Testing attachment, section 2b2.3

- This measure is specified for both group and individual clinicians, which by NQF criteria must be tested separately. The testing was performed with these two groups pooled together, and hence is not tested to specifications.
- The results of empirical validity was a moderate positive correlation at the registry level (0.57). However, result was a weak positive correlation at the claim level (0.22). The developer noted PQRS 117 is only existing measure they could correlate with 0086.
- For face validity testing results from 2013, the mean rating was 4.56 and 87.50% of TEP either agreed or strongly agreed that this measure can accuarately distinguid good and poor quality.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

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

    ☐ **Yes**

    ☐ **No**

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

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

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

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

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

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

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

- The developer previously conducted face validity on this measure and performed convergent validity testing.
- The convergent validity testing had weak positive correlation results at the claims level and moderate positive correlation results at the registry level. This isn't strong evidence for the validity of the claims.

ADDITIONAL RECOMMENDATIONS

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

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

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

- I have a question whether the data element of optic nerve head evaluation includes documentation of the stability of current situation compared to prior, or documented change, a number in isolation (cup to disc ratio), or an evaluation of ancillary testing (ie OCT of the nerve fiber layer) This was also mentioned in the PCPI internal review. There may be some inconsistency regarding the definition of primary open angle glaucoma in this measure, as the ICD-10 codes include those for low-pressure (normal tension) glaucoma which are not typically included in my mind as part of the definition for POAG.
- The data elements are clearly defined and the codes are provided. There are no concerns.
- I don't know how multiple evaluations are handled via this measure. Meaning that if the patient presented 10 times in the year and you evaluated the optic nerve only once, is that factored into the measure. WE need to determine the frequency of this evaluation.
- Insufficient, as the measure wasn't tested to the specifications as noted in the submission document. The measure does not appear to capture unspecified codes for example: H40.1294 (Low-tension glaucoma, unspecified eye, indeterminate stage or H40.1193 Primary open-angle glaucoma, unspecified eye, severe

stage. Understanding the goal is to code to the highest level of specificity however there is a potential set of the population that could be overlooked without including these codes.

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

- No
- No
- No
- Insufficient, as the measure wasn't tested to the specifications as noted in the submission document. The measure does not appear to capture unspecified codes for example: H40.1294 (Low-tension glaucoma, unspecified eye, indeterminate stage or H40.1193 Primary open-angle glaucoma, unspecified eye, severe stage. Understanding the goal is to code to the highest level of specificity however there is a potential set of the population that could be overlooked without including these codes.

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

- Inconsistent claims and registry validity testing. However, it is possible that because the measure was compared with diabetic retinopathy documentation, and some specialty care is fragmented between retinal and glaucoma specialists, that different doctors may be performing/documenting different aspects of the examination. The composition of specialty doctors may possibly be different between the claims and the registry population
- The measure was not tested separately for groups and individuals., although this was unchanged from previous testing.
- No
- No, it appears the developer tested the measure for with claims and registry data and yielded appropriate results.

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

- Registry- Looking at the performance percentiles without outliers, 10% of the data falls below a performance score of 0.97. Claims- Looking at the performance percentiles without outliers, 100% of the data is equal to a performance score of 1.0. Suggests performance is high.
- The developer did not test for missing data but I do not think it constitutes a threat to validity.
- No.
- Some of the testing documents are i.e. testing for group and individual clinicians was pooled and the empirical validity results were not high, I would agree with the low overall ranking for validity

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

- n/a
- The exclusions are consistent and there are no risk adjustments.
- There was no case adjustment proposed or added from my review of this measure.
- 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.

- The measure is constructed using claims and registry data. The eMeasure version is #0086e and will be evaluated separately, data for the eMeasure will be accessed via electronic sources.
- All data elements are in defined elements in a combination of electronic sources.
- Developer notes they have included some proprietary coding in the measure for convenience, and fees may be required to use those.

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

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

- Normally documentation of the optic nerve head is performed in outpatient charting, and is typically well-defined in an EHR. The nerve can be described in terms of disc size, cup-to-disc ratio, presence of notching, presence of peripapillary atrophy, or disc hemorrhages. There may be some variability in what constitutes an actual diagnosis of Glaucoma as more diagnostic tools become available (ie will pre-perimetric glaucoma be considered 'glaucoma'?)
- There are no concerns with feasibility.
- All required elements are generated routinely.
- Yes, the tools are available. However, there continues to be a barrier in the utilization of CPT II codes and thus is a challenge to collect the 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

- Merit-based Incentive Payment System (MIPS) – results will be available in Physician Compare starting this year
- IRIS™ Registry (Intelligent Research in Sight) – American Academy of Ophthalmology comprehensive eye disease and condition registry
- MORE Registry (Measures and Outcomes Registry for Eyecare) - qualified clinical data registry (QCDR) sponsored by the American Optometric Association

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

Feedback on the measure by those being measured or others

- The developer noted that they maintain several pathways for feedback on the measure including the use of topic-specific technical expert panels during the measure development and during the annual maintenance process, as well as feedback via an online public comment and an email-based process set up to receive measure inquiries from implementers.
- Developers noted that during the development of the measure, clinicians requested for definition of "what is considered an adequate examination of the optic nerve." The TEP clarified clarified that ultimately the physicians will use their best available tools to perform the optic nerve evaluation.
- The public also highlighted the limitations of the age range covered by this measure; patient population younger than 65 might not be enrolled in Medicare programs, causing a limitation in implementation and tracking of the measure. The developer also clarified concerns about age limitation for individuals not enrolled in Medicare. They asserted that the measure can be adapted by physicians, payers and other interested groups who want to use this measure as a metric for quality improvement.

Additional 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**

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

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

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

Improvement results

- The measure developer reported a 5% decrease in performance rate from 2013 to 2017.
- The goal of this measure is to promote examination and documentation of the structure and function of the optic nerve and to monitor and detect disease progression among POAG patients.

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 noted no unexpected findings, positive or negative to have been recorded during implementation.

Potential harms

- The developer noted no potential harms to have been documented from use of this measure.

Additional Feedback: N/A

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do you agree with the measure developer that there are no unintended consequences associated with the measure?

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

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

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

- MIPS, MORE, IRIS registry. PCPI has a forum for this process and is detailed in the submission
- Feedback has been provided for those being measured.
- I'm not clear if those people are provided feedback on missing the measure.
- Data available and feedback is obtain by the developer.

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

- The performance rate is high in some of the registry and claims data, but lower in the optometry registry data. This disparity may warrant further investigation.
- There do not appear to be any unintended consequencs.
- There are no unintended harms of this evaluation that I can think of.
- The evaluation of the optic nerve per the research is a good way to monitor and detect disease progression among POAG patients, hence beneficial. There is concern that there is a decline in performance which could support the need to continue to evaluate the measure.

Criterion 5: Related and Competing Measures

Related or competing measures

0563 : Primary Open-Angle Glaucoma: Reduction of Intraocular Pressure by 15% or Documentation of a Plan of Care

Harmonization

NQF #0563 evaluates reduction in intraocular pressure for patients with glaucoma. NQF #0086 measures the evaluation of the optic nerve to establish glaucoma disease status and presence of optic nerve damage. AAO recommends a 20-30% reduction in intraocular pressure from baseline for patients with tension glaucoma. NQF 0563 measures this reduction in IOP from baseline level. and NQF 0086 measures optic nerve healthand monitors, detects and prevents disease progression of Primary Open Angle Glaucoma. 0563 does not evaluate patients who have normal or low tension glaucoma, and 0086 does include these patients.

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

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

- Another glaucoma measure evaluates lowering intraocular pressure from baseline state. These two measures are complimentary but adequately different such that they should not likely be combined. There are some exclusions in the other measure based on definition of primary open angle glaucoma that are appropriate.
- Measure 0563 is an outcome measure and does not compete.
- Not applicable.
- Yes, 0563 which appears to be harmonized with the measure and provides more outcomes related results

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/12/2019

- **No NQF Members have submitted support/non-support choices as of this date.**

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

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

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

2019_POAG_0086_NQF_evidence_attachment_v7.1_FINAL-636911080547577825.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

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0086

Measure Title: Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation

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

Date of Submission: 4/9/2019

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*):

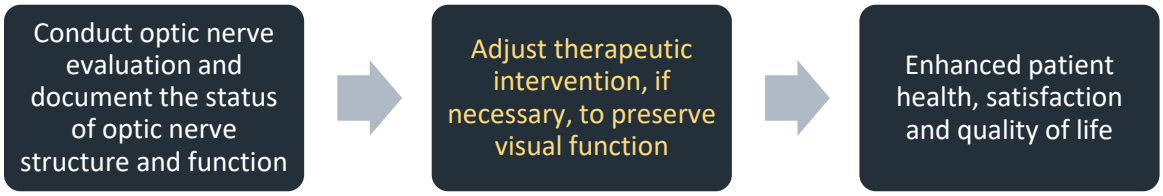
☒ Process: Optic nerve evaluation for patients diagnosed with primary open-angle glaucoma

☐ 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.



Changes in the optic nerve are one of two characteristics which currently define progression and thus worsening of glaucoma (the other characteristic is visual field). A stable optic nerve is one of the goals of managing patients with primary open-angle glaucoma (1). Follow-up evaluation and documentation of the optic nerve head and retinal nerve fiber layer, among patients with primary-open angle glaucoma, provide valuable structural information about glaucomatous optic nerve damage, and informs the therapeutic goals to preserve visual function. For example, indications for therapy adjustment among primary open-angle glaucoma patients include progressive optic nerve damage despite achieving target intraocular pressure.

People with vision loss are more likely to report depression, diabetes, hearing impairment, stroke, falls, cognitive decline, and premature death (2). Decreased ability to see often leads to the inability to drive, read, keep accounts, and travel in unfamiliar places, thus substantially compromising quality of life (2). Furthermore, in a 2005 Survey of Public Knowledge, Attitudes, and Practices Related to Eye Health and Disease, 71 percent of respondents answered that loss of eyesight would have the greatest impact on their daily life (3). Consequently, this measured process leads to a desired health outcome of preservation of visual function and ultimately the maintenance of quality of life for the patient.

For high tension glaucoma patients, when initiating therapy, the clinician sets a target range of controlled IOP based on the pretreatment pressure and the presence of optic nerve damage. According to the AAO Glaucoma Preferred Practice Pattern, lowering the pretreatment IOP by 25% or more has been shown to inhibit progression of POAG to preserve visual function.

(1) American Academy of Ophthalmology's Glaucoma Preferred Practice Pattern Panel. (2015). Preferred Practice Pattern Guidelines - Primary Open-Angle Glaucoma. San Francisco, CA: American Academy of Ophthalmology. Retrieved from https://www.aao.org/preferred-practice-pattern/primary-open-angle-glaucoma-ppp-2015.

(2) Centers for Disease Control and Prevention – Vision Health Initiative. (2015). Why is Vision Loss a Public Health Problem? Retrieved from https://www.cdc.gov/visionhealth/basic_information/vision_loss.htm

(3) National Eye Institute & Lions Clubs International Foundation. (2007). 2005 survey of public knowledge, attitudes, and practices related to eye health and disease. Bethesda, MD: National Eye Institute. Retrieved from: http://www.nei.nih.gov/nehep/kap/.

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

Not applicable

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

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

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

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses

explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	American Academy of Ophthalmology Glaucoma Panel. Preferred Practice Pattern Guidelines. Primary Open-Angle Glaucoma. San Francisco, CA: American Academy of Ophthalmology; 2010. Available at www.aao.org/ppp. *Note that while PCPI has made nominal updates to this evidence attachment since the last NQF submission, the underlying evidence and intent of the measure have not changed. Updates were made to capture the current language in the most recent guideline, in support of the measure.* • Title: Preferred Practice Pattern. Primary Open-Angle Glaucoma. American Academy of Ophthalmology. • Author: Bruce E. Prum, Jr., Lisa F. Rosenberg, Steven J. Gedde, Steven L. Mansberger, Joshua D. Stein, Sayoko E. Moroi Leon W. Herndon, Jr., Michele C. Lim, Ruth D. Williams • Date: September 18, 2015 • Citation, including page number: American Academy of Ophthalmology's Glaucoma Preferred Practice Pattern Panel. Preferred Practice Pattern Guidelines. Primary Open-Angle Glaucoma. San Francisco, CA: American Academy of Ophthalmology; 2015. P58-60, P76. • URL: https://www.aao.org/preferred-practice-pattern/primary-open-angle-glaucoma-ppp-2015

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.	Ophthalmic Evaluation recommendation; page 12. In completing the elements in the comprehensive adult medical eye evaluation, the ophthalmic evaluation specifically focuses on the following elements: History [A:III] Visual acuity measurement [A:III] Pupil examination [B:II] Anterior segment examination [A:III] Intraocular pressure measurement [A:I] Gonioscopy [A:III] **Optic nerve head and retinal nerve fiber layer examination [A:III]** Fundus examination [A:III) *Note that since the last submission to NQF, the AAO has updated its Preferred Practice Patterns to reflect methodology for grading the strength of evidence and recommendations, from SIGN and GRADE groupings.* The optic nerve should be carefully examined for signs of glaucoma damage, and its appearance should be serially documented (I+, moderate quality, strong recommendation).
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Strength of evidence rating: Level III I+: Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias Moderate Quality Body of Evidence: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

Provide all other grades and definitions from the evidence grading system	Strength of Evidence Ratings: Level I: Randomized controlled trial or meta-analyses Level II: Controlled trials, cohort, or case-control studies Level III: Descriptive studies or case reports To rate individual studies, a scale based on SIGN is used. The definitions 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++ 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 • II- Case-control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal • III Nonanalytic studies (e.g., case reports, case series) Recommendations for care are formed based on the body of the evidence. The quality ratings for the body of evidence are defined by GRADE as follows: • Good quality: Further research is very unlikely to change our confidence in the estimate of effect • 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
Grade assigned to the **recommendation** with definition of the grade	Care process rating – A Care process rating: Level of Importance: Level A, defined as most important Strong Recommendation: Used when the desirable effects of an intervention clearly outweigh the undesirable effects
Provide all other grades and definitions from the recommendation grading system	Recommendations of Care Ratings Care Process Ratings: Level A: Most important to the care process Level B: Moderately important to the care process Level C: Relevant but not critical to the care process 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 • 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

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The evidence cited in support of the guideline recommendation is comprised mainly of descriptive studies including 9 prospective clinical case series. The evidence also includes data from analytical studies including 1 prospective case-control and 1 prospective cohort study. Also cited were 2 randomized control trials including the Early Manifest Glaucoma Trial and the European Glaucoma Prevention Study. While the current guideline does not provide an analysis of the quantity and quality of the evidence supporting this measure, we analyzed the studies corroborating the guideline recommendation and they are as follows: Quantity: 38 studies were cited in support of the guideline recommendation. Quality: 32 of the studies in support of this guideline recommendation were observational study designs, including 7 descriptive studies, 23 cohort studies, and 2 case/control studies. 6 were randomized controlled trials, including the Early Manifest Glaucoma Trial Group, Collaborative Normal-Tension Glaucoma Study Group, and the European Glaucoma Prevention Study. The findings of these studies highlighted the importance of careful examination of the optic nerve head, including checking for the presence of a disc hemorrhage which is an important biomarker for glaucoma damage.
Estimates of benefit and consistency across studies	The guideline does not provide a quantitative estimate of benefit across studies for the assessment of the optic nerve to distinguish disease progression in patients with primary open-angle glaucoma. The studies cited are summarized to indicate that examination of the optic nerve head and retinal nerve fiber layer provides valuable structural information about optic nerve damage resulting from glaucoma. The guideline states that visible structural alterations of the optic nerve head or retinal nerve fiber layer frequently occur before visual field defects can be detected and careful study of the optic disc neural rim for small hemorrhages is important, since these hemorrhages often precede visual field loss and further optic nerve damage in patients with glaucoma. (AAO, 2010) Estimates of benefit and consistency across studies supporting the measure were not directly addressed in the AAO guideline. However, the AAO recommendation in support of this measure received a “Strong Recommendation” which indicates that the desirable effects of evaluating the optic nerve head among patients with POAG outweigh the undesirable effects. AAO found that careful study of the optic disc neural rim for small hemorrhages is important because these hemorrhages sometimes herald focal disc damage and visual field loss, and they may signify ongoing optic nerve damage in patients with glaucoma (AAO, 2015, P60). Tracking this optic nerve damage allows the clinician to adjust the therapeutic intervention if necessary, to ultimately preserve visual function.

What harms were identified?	The guideline does not delineate harms resulting from an evaluation of the optic nerve. The guideline does indicate the expected benefits of assessing the optic nerve, which include: • Preventing further optic nerve damage by estimating an appropriate intraocular pressure (IOP) target level • Maintaining a patient's visual function by initiating appropriate therapeutic interventions to maintain IOP at or below IOP target level The current guideline does not delineate harms.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	There were no significant updates to the body of evidence conducted since the systematic review that would have an impact on the conclusions about the importance of optic nerve head examination, as changes in the optic nerve is one of two characteristics which currently define progression and thus worsening of glaucoma. There were no significant updates to the body of evidence conducted since the systematic review that would have an impact on the conclusions about the importance of optic nerve head examination.

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.

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.

Glaucoma is a group of diseases that damage the eye's optic nerve and can result in vision loss and blindness. In 2011, 2.71 million persons in the U.S. had primary open-angle glaucoma (POAG) and in 2050, an estimated 7.32 million persons will have POAG (1). Furthermore, a 2006 study estimated that the total financial burden of major visual disorders among U.S. residents aged 40 years or older was $35.4 billion in 2004: $16.2 billion in direct medical costs, $11.1 billion in other direct costs, and $8 billion in productivity losses. Of the direct medical costs, approximately $2.9 billion was attributable to glaucoma (2). It is imperative that evidence-based care be delivered to all glaucoma patients.

According to the most recent guidelines, changes in the optic nerve are one of two characteristics which currently define progression and thus worsening of glaucoma disease status (the other characteristic is visual

field). Examination of the optic nerve head (ONH) and retinal nerve fiber layer (RNFL) provides valuable structural information about glaucomatous optic nerve damage. Visible structural alterations of the ONH or RNFL and development of parapapillary choroidal atrophy in early glaucoma may precede the onset of visual field defects. Careful study of the optic disc neural rim for small hemorrhages is important because these hemorrhages sometimes herald focal disc damage and visual field loss, and they may signify ongoing optic nerve damage in patients with glaucoma (3). Despite evidence emphasizing the value of an optic nerve evaluation, there is a gap in documentation patterns of the optic nerve for both initial and follow-up care.

This measure is intended to promote examination and documentation of the structure and function of the optic nerve, and to monitor and detect disease progression among POAG patients. This measure should lead to the desired health outcome of preservation of one's visual function and ultimately the maintenance of quality of life for the patient.

1. Vajaranant, T. S., Wu, S., Torres, M., & Varma, R. (2012). The Changing Face of Primary Open-Angle Glaucoma in the United States: Demographic and Geographic Changes From 2011 to 2050. American Journal of Ophthalmology, 154(2). doi:10.1016/j.ajo.2012.02.024

2. Rein, D. B., Zhang, P., & Wirth, K. (2006). The Economic Burden of Major Adult Visual Disorders in the United States. Archives of Ophthalmology, 124(12), 1754-1760. doi:10.1001/archopht.124.12.1754

3. Prum, B. E., Rosenberg, L. F., Gedde, S. J., Mansberger, S. L., Stein, J. D., Moroi, S. E., . . . Williams, R. D. (2015). Primary Open-Angle Glaucoma Preferred Practice Pattern® Guidelines. Ophthalmology, 123(1). doi:10.1016/j.ophtha.2015.10.053

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.*

POAG 0086 Registry

2016 Registry data from the PQRS program was provided to the PCPI by CMS for the purposes of testing the measure.

The data are analyzed for the time period January 2016 through December 2016 and include 152,753 quality events. The mean performance rate is 0.94, the standard deviation is 0.15, the minimum is 0.04, the maximum is 1.00, and the interquartile range is 0.03 (1.00 – 0.97). Performance Scores by Decile: (1st,0.81; 2nd,0.94; 3rd,0.99; 4th,1.00; 5th,1.00; 6th, 1.00; 7th,1.00; 8th,1.00; 9th,1.00; 10th,1.00)

Historical PQRS data from the PQRS Experience Report does not differentiate between EHR, Claims, and Registry average performance rates. Performance scores over time are for 2013: 0.95, 2014: 0.94, 2015: 0.91

POAG 0086 Claims

2016 Claims data from the PQRS program was provided to the PCPI by CMS for the purposes of testing the measure.

The data are analyzed for the time period January 2016 through December 2016 and include 499,801 quality events. The mean performance rate is 0.97, the standard deviation is 0.11, the minimum is 0.01, the maximum is 1.00, and the interquartile range is 0.00 (1.00 – 1.00). Performance Scores by Decile: (1st,0.96; 2nd,1.00; 3rd,1.00; 4th,1.00; 5th,1.00; 6th, 1.00; 7th,1.00; 8th,1.00; 9th,1.00; 10th,1.00)

Historical PQRS data from the PQRS Experience Report does not differentiate between EHR, Claims, and Registry average performance rates. Performance scores over time are for 2013: 0.95, 2014: 0.94, 2015: 0.91

CMS published the following data in its 2017 Quality Payment Program Experience Report (1) and 2016 PQRS Reporting Experience Report (2), for Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation. Experience report data does not differentiate among EHR, Registry, and Claims average performance rates. It is important to note that PQRS was a voluntary reporting program which is reflected in the reporting rate among

those eligible to report on this measure. In some cases, the reporting rate was as low as 37% for this measure. We also know that participation in the program overall was suboptimal, with 72% of eligible professionals using any method to participate in PQRS, in 2016. The performance scores listed below are not consistently derived from a nationally representative sample.

Year	Modality	Average Performance Rate
2017	QPP	90.17%
2016	PQRS	91.6%
2015	PQRS	91%
2014	PQRS	94%
2013	PQRS	95.4%

Year	Modality	Reporting Rate (percentage of those eligible to report on measure)
2017	QPP	85.06%
2016	PQRS	36.9%
2015	PQRS	43.7%
2014	PQRS	44.5%
2013	PQRS	38.1%

(1) 2017 Quality Payment Program Reporting Experience. Available at: https://qpp-cm-prod-content.s3.amazonaws.com/uploads/492/2017%20QPP%20Experience%20Report%20Appendix.zip.

(2) 2016 Reporting Experience Including Trends (2007-2016), Physician Quality Reporting System. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Downloads/2016-Appendix.xlsx.

We also received average performance rates from the American Optometric Association (AOA) Measures and Outcomes Registry for Eyecare (MORE) Registry/QCDR, for this measure:

Year	Modality	Average Performance Rate
2018	QCDR	75%
2017	QCDR	53%

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.

In an analysis of a physician-led, team-based care model for treating glaucoma, records of 591 patients with newly diagnosed glaucoma were assessed retrospectively, amongst two three-year periods, for the completion of 9 American Academy of Ophthalmology (AAO) Preferred Practice Pattern (PPP) recommended metrics. The primary outcome was the percent of patients with completion of each examination component. The study looked at testing completed within three visits of initial glaucoma diagnosis at Mayo Clinic Health System and found that ophthalmologists had poor adherence to measuring the cup to disk ratio at 79.6% from 2005 to 2007 and 83.6% in 2008-2010. (1)

In a sample of 300 charts (3650 visits) which included optic disc examination results by clinical, photographic, and imaging techniques, physicians varied dramatically in their adherence to the American Academy of Ophthalmology Preferred Practice Pattern on open-angle glaucoma, performing disc evaluations and imaging on 90 percent of open-angle glaucoma patients. The study also cited that the annualized rate for recording the

cup-to-disc status was once yearly or more in 66% of patients. The study concluded that physician adherence to practice guidelines varied substantially, and therefore scoring systems for physician behavior have promise in measuring outcome improvements related to better care. (2)

(1) Winkler, N., Damento, G., Khanna, S., Hodge, D. and Khanna, C. (2017). Analysis of a Physician-led, Team-based Care Model for the Treatment of Glaucoma. Journal of Glaucoma, 26(8), pp.702-707.

(2) Quigley, H. A., Friedman, D. S., & Hahn, S. R. (2007). Evaluation of Practice Patterns for the Care of Open-angle Glaucoma Compared with Claims Data. Ophthalmology, 114(9), 1599-1606. doi:10.1016/j.ophtha.2007.03.042

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 this measure is included in a federal reporting program, the program does not provide disparities data to analyze and report. In Section 1b.5 below, we provide disparities data reported in the literature.

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 retrospective longitudinal cohort study looked at 21,766 adults with newly diagnosed open-angle glaucoma (OAG) between 2007-2011 and enrolled in either Medicaid or a large U.S. managed care network. The study concluded that Medicaid beneficiaries with OAG received substantially less glaucoma testing compared to those who had commercial health insurance. Specifically, the proportions of beneficiaries with commercial health insurance, with newly diagnosed OAG, who underwent visual field testing, fundus photography, or other ocular imaging were 63%, 22%, and 54% respectively. On the other hand, the proportions of Medicaid beneficiaries to receive those same tests to monitor OAG were 35%, 19%, and 30% respectively. Compared to those with commercial health insurance, Medicaid recipients were 234% more likely to not receive any glaucoma testing in the 15 months following initial diagnosis (OR=3.34, CI:3.07-3.63). (1) At the same time, Black Americans age 40 and older are at the highest risk of developing open-angle glaucoma, compared with people of other races (2) and comprise 21% of the Medicaid population (3), indicating racial disparities amongst those who receive glaucoma testing following initial diagnosis.

A case-control study of individuals with glaucoma or suspected glaucoma, at a county hospital, found racial disparities among those who adhered to consistent clinician follow-up visits. Of the "cases" (defined in the study as inconsistent follow-up), 27.6%, 40.8%, and 7.9% were Black, Latino, and White, respectively. (4)

A retrospective cohort study of glaucoma patients and individuals who had cupping of the optic disc, who were enrolled at a large managed care organization, found that women were 24% less likely to undergo treatment than men (odds ratio, 0.76; 95% confidence interval, 0.71-0.80). Note that the logistic regression model adjusted for glaucoma status, age, region, clinician seen at initial visit, and index date. (5)

(1) Elam, A. R., Andrews, C., Musch, D. C., Lee, P. P., & Stein, J. D. (2017). Large Disparities in Receipt of Glaucoma Care between Enrollees in Medicaid and Those with Commercial Health Insurance. Ophthalmology, 124(10), 1442-1448. doi:10.1016/j.ophtha.2017.05.003.

(2) NIH National Eye Institute. Glaucoma, Open-angle. (2010). Retrieved from https://nei.nih.gov/eyedata/glaucoma.

(3) Kaiser Family Foundation. Medicaid Enrollment by Race/Ethnicity. (2017, December 12). Retrieved from https://www.kff.org/state-category/medicaid-chip/medicaid-beneficiaries/.

(4) Murakami Y, Lee BW, Duncan M, et al. (2011). Racial and Ethnic Disparities in Adherence to Glaucoma Follow-up Visits in a County Hospital Population. Arch Ophthalmol, 129(7), 872–878. doi:10.1001/archophthalmol.2011.163.

(5) Friedman, D. S., Nordstrom, B., Mozaffari, E., & Quigley, H. A. (2005). Variations in Treatment among Adult-Onset Open-Angle Glaucoma Patients. Ophthalmology, 112(9), 1494-1499. doi:10.1016/j.ophtha.2005.02.010.

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):*

Ears, Nose, Throat (ENT), Eye Care

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

Health and Functional Status : Change

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.)*

The measure specifications are included with this form. Additional measure details may be found at http://www.thepcpi.org/?page=PCPIMeasures

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:** NQF0086_I9toI10_conversion.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. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Supporting guidelines and coding value sets included in the measure are reviewed on an annual basis. This annual review has resulted in the removal of coding related to 'unspecified eye,' as these codes were determined by clinical experts to have low yield and to represent poor documentation practices.

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

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

Patients who have an optic nerve head evaluation during one or more office visits within 12 months

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

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

Time Period for Data Collection: At least once during the measurement period

Report CPT Category II Code, 2027F: Optic nerve head evaluation performed

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

All patients aged 18 years and older with a diagnosis of primary open-angle glaucoma

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

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

Time Period for Data Collection: 12 consecutive months

Patients aged >= 18 years on date of encounter

AND

Diagnosis for primary open-angle glaucoma (ICD-10-CM): H40.10X0, H40.10X1, H40.10X2, H40.10X3, H40.10X4, H40.1110, H40.1111, H40.1112, H40.1113, H40.1114, H40.1120, H40.1121, H40.1122, H40.1123, H40.1124, H40.1130, H40.1131, H40.1132, H40.1133, H40.1134, H40.1210, H40.1211, H40.1212, H40.1213, H40.1214, H40.1220, H40.1221, H40.1222, H40.1223, H40.1224, H40.1230, H40.1231, H40.1232, H40.1233, H40.1234, H40.151, H40.152, H40.153

AND

Patient encounter during the performance period (CPT): 92002, 92004, 92012, 92014, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, 99241, 99242, 99243, 99244, 99245, 99304, 99305, 99306, 99307, 99308, 99309, 99310, 99324, 99325, 99326, 99327, 99328, 99334, 99335, 99336, 99337

WITHOUT

Telehealth Modifier: GQ, GT, 95, POS 02

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

Denominator Exceptions:

Documentation of medical reason(s) for not performing an optic nerve head evaluation

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.)*

Time Period for Data Collection: During the encounter within the 12-month period

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. For measure Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation, exceptions may include medical reason(s) for not performing an optic nerve head evaluation. Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Append a modifier to CPT Category II Code, 2027F-1P: Documentation of medical reason(s) for not performing an optic nerve head evaluation

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.)*

Consistent with CMS' Measures Management System Blueprint and national recommendations put forth by the IOM (now NASEM) and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer.

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.)*

To calculate performance rates:

1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).

2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.
3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator.
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: medical reason(s) for not performing an optic nerve head evaluation]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. -- Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents a quality failure.

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

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

Not applicable. The measure is not based on a sample.

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

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

Not applicable. The measure is not based on a survey.

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

If other, please describe in S.18.

Claims, Registry 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.

Not applicable.

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)*

Other, Outpatient Services, Post-Acute Care

If other: Domiciliary

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

Not applicable. The measure is not a composite.

2. Validity – See attached Measure Testing Submission Form

v2_0086rc_nqf_testing-attachment_7.1-636849651617196441.docx,0086rc_MAR282019_nqf_testing-attachment_7.1_Final.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

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

Measure Number (*if previously endorsed*)**:** 0086
Measure Title: Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation
Date of Submission: 3/28/2019

Type of Measure:

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

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

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

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

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

Previously Submitted 2013 Data

Registry

The data source is the Centers for Medicare & Medicaid Services PQRS GPRO database.

Claims

The data source is the Centers for Medicare & Medicaid Services PQRS administrative claims database.

Current Testing Data

The data source is claims and registry data from the Physician Quality Reporting System (PQRS), provided by the Centers for Medicare & Medicaid Services (CMS).

To participate in the PQRS program, Eligible Professionals and Group practices submit performance data such as number of eligible instances (denominator), instances of quality service performed (numerator), number of

performance exclusions, reporting rates, and performance rates—in a file format specified by CMS. Data is then summarized at the practice level and includes both EPs participating individually as well as group practices participating through GPRO.

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

Previously Submitted 2013 Data

The claims and registry data are for the time period January 2013 – December 2013 and cover the entire United States.

Current Testing Data

The claims and registry data are for the time period January 2016 through December 2016 and cover the entire United States. Given the required conversion to ICD-10 in late 2015, the testing was completed on the ICD-10 specified measure.

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*)

Previously Submitted 2013 Data

Registry

For this measure, the minimum number required to be included is 10 events. Given the structure of the PQRS program, a physician may choose to submit or not submit to PQRS on any given claim. Since these data contain results on a large number of physicians, limiting the reliability analysis to only those physicians who are participating in the program will eliminate the bias introduced by the inclusion of from physicians who are in the data but are not submitting claims to PQRS.

The total number of physicians reporting on this measure is 1,824. Of those, 1,359 physicians had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. For this measure, 74.5 percent of physicians are included in the analysis, and the average number of quality reporting events is 121.8 for a total of 165,527 events. The range of quality reporting events for 1,359 physicians included is from 2,773 to 10. The average number of quality reporting events for the remaining 25.49 percent of physicians who aren't included is 3.7.

Claims

For this measure, the minimum number required to be included is 10 events. Given the structure of the PQRS program, a physician may choose to submit or not submit to PQRS on any given claim. Since these data contain results on a large number of physicians, limiting the reliability analysis to only those physicians who are participating in the program will eliminate the bias introduced by the inclusion of from physicians who are in the data but are not submitting claims to PQRS.

The total number of physicians reporting on this measure is 44,998. Of those, 9,616 physicians had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. For this measure, 21.4 percent of physicians are included in the analysis, and the average number of quality reporting events is 92.7 for a total of 891,018 events. The range of quality reporting events for 9,616 physicians included is from 1,508 to 10. The average number of quality reporting events for the remaining 78.6 percent of physicians who aren't included is 0.6.

Current Testing Data

Registry

We received data from 1,591 providers reporting on this measure through the registry reporting option for CMS's PQRS in 2016. This dataset reflects a combination of individual provider data and group data and our analysis of the data as a whole is reflected throughout this submission. 1,591 providers had all the required data elements and at least one quality reporting event for a total of 152,753 quality events. For this measure, 100 percent of providers are included in the analysis, and the average number of quality reporting events is 96. The range of quality reporting events for 1,591 providers included is from 1 to 2,712.

Claims

We received data from 13,685 providers reporting on this measure through the claims reporting option for CMS's PQRS in 2016. This dataset reflects a combination of individual provider data and group data and our analysis of the data as a whole is reflected throughout this submission. Of those, 13,377 providers had all the required data elements and at least one quality reporting event for a total of 499,801 quality events. The remainder of providers had 0 quality reporting events after accounting for exceptions. For this measure, 98.0 percent of providers are included in the analysis, and the average number of quality reporting events is 37. The range of quality reporting events for 13,377 providers included is from 1 to 1,368.

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*)

Previously Submitted 2013 data

Registry

There were 165,528 patients included in this testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

Claims

There were 891,018 patients included in this testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

Current Testing Data

Registry

There were 152,753 quality events included in this reliability testing and analysis. These were the quality events that were associated with providers who had all the required data elements.

Claims

There were 499,801 quality events included in this reliability testing and analysis. These were the quality events that were associated with providers who had all the required data elements.

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.

Previously Submitted 2013 Data

Registry and Claims

The same data sample from each data source was used for the respective reliability testing, performance testing, and exceptions analysis.

Face Validity

After the measure was fully specified, an expert panel of 16 members were asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Current Testing Data

Registry and Claims

The same data samples were used for reliability testing and exceptions analysis.

After conducting a thorough evaluation of available and relevant PQRS data we selected Diabetes: Eye Exam (PQRS #117) for empirical validity correlation.

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.

Previously Submitted 2013 Data

Registry and Claims

This was not captured as part of the testing.

Current Testing Data

Registry and Claims

Patient-level socio-demographic (SDS) variables were not captured as part of the testing as that information was not provided in the CMS data used for analysis.

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*)

Previously Submitted 2013 Data

Data 2 (Claims) and Data 3 (Registry)

Reliability of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

Current Testing Data:

Data from the both the previous testing data and the current testing data samples were tested using the same reliability testing method.

Reliability of the computed measure score was measured as 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 provider performance and the noise is the total variability in measured performance.

Reliability at the level of the specific provider is given by:

Reliability = Variance (provider-to-provider) / [Variance (provider-to-provider) + Variance (provider-specific-error]

Reliability is the ratio of the provider-to-provider variance divided by the sum of the provider-to-provider variance plus the error variance specific to a provider.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the provider performance score is a binomial random variable conditional on the provider'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.

Reliability is evaluated by averaging over provider specific reliabilities for all providers that have at least 1 quality reporting event for the measure.

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in provider performance. A reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. [1]

1. Adams JL, Mehrotra A, McGlynn EA, Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on February 24, 2012.)

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*)

Previously Submitted 2013 Data

Registry

For this measure, the reliability at the minimum level of quality reporting events (10) was 0.72. The average number of quality reporting events for physicians included is 121.8. The reliability at the average number of quality reporting events was 0.97.

Claims

For this measure, the reliability at the minimum level of quality reporting events (10) was 0.86. The average number of quality reporting events for physicians included is 92.7. The reliability at the average number of quality reporting events was 0.98.

Current Testing Data

Registry

The average reliability including providers with at least one quality reporting event is 0.95. Reliability was also evaluated at the 10th, 25th, 50th, 75th, and 90th cut points:

	1+ Events
Percentile	Value
10th	0.86
25th	0.98
50th	1.00
75th	1.00
90th	1.00

Claims

The average reliability including providers with at least one quality reporting event is 0.96. Reliability was also evaluated at the 10th, 25th, 50th, 75th, and 90th cut points:

	1+ Events
Percentile	Value
10th	0.95
25th	1.00
50th	1.00
75th	1.00
90th	1.00

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?*)

Previously Submitted 2013 Data

Registry

This measure has moderate reliability when evaluated at the minimum level of quality reporting events and high reliability at the average number of quality events.

Claims

This measure has high reliability when evaluated at the minimum level of quality reporting events and high reliability at the average number of quality events.

Current Testing Data

Registry

This measure has high reliability when including providers with at least one quality reporting event.

Claims

This measure has high reliability when including providers with at least one quality reporting event.

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)*

Previously Submitted 2013 Face Validity Data

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 5= Strongly Agree

Current Testing Data – empirical validity correlation testing

For this measure, the PCPI has conducted review and updates to the measure specifications, which satisfy the NQF's ICD-10 Conversion requirements. We are providing the information below to support the three requirements:

- NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM
  Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.
- NQF ICD-10-CM Requirement 2: Coding Table
  See attachment in S.2b
- NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes
  The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have an RHIA-credentialed professional on our staff who reviews all ICD-10 coding. For measures included in CMS' Quality Payment Program (QPP), the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comments received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

Empirical validity correlation testing

Diabetes: Eye Exam (PQRS #117) was chosen as a suitable candidate for correlation analysis due to the similarities in patient population and domain. We hypothesize that there exists a positive association of scores between providers who performed an optic nerve head evaluation on patients with primary open angle glaucoma at least once within the 12-month measurement period and those who performed an eye exam (retinal) on patients with diabetes (types 1 and 2).

Providers included in the analysis had at least one quality reporting event and were cleaned in the same process as the PQRS dataset.

Datasets were reviewed to identify shared providers based on NPI and TIN identifiers. Comparing performance scores of those shared providers, the empirical analysis uses regression with dataset 1 as the outcome and dataset 2 as the predictor. Results identify the multiple R value (the correlation coefficient) and P-value of the regression variables to assess the association between performance scores of these shared providers.

We use the following guidance to describe correlation[1]:

Correlation	Interpretation
0.80 – 1.00	Very Strong
0.60 – 0.79	Strong
0.40 – 0.59	Moderate
0.20 - 0.39	Weak
0 – 0.19	Very Weak

1. "11. Correlation and Regression." *The BMJ*, 21 March 2019, https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression/ .

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

Previously Submitted 2013 Data

Face Validity

Our expert panel included 16 members. The list of expert panel members is as follows:

Gregory Kwasny, MD

David J. Forster, MD

John McAllister, MD

David B. Glasser, MD

Michael Repka, MD

Trexler M. Topping, MD

Jeffrey P. Edelstein, MD

Sonya Shah, MD

John M. Haley, MD

George Williams, MD

Joseph LoCascio, MD

Cynthia Mattox, MD

Daniel Briceland, MD

Kristin Carter, MD

Craig Kliger, MD

Bradley Fouraker, MD

Current Testing Data – empirical validity correlation testing

Registry

Data from the PQRS program were used to perform the correlation analysis for this measure. Data comes from the registry versions of Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation (PQRS #012) and Diabetes: Eye Exam (PQRS #117).

Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation (PQRS #012) demonstrates positive correlation with Diabetes: Eye Exam (PQRS #117).

PQRS #117

Coefficient of correlation = 0.57

Alpha level = 0.05

P-value < 0.001

Number of shared providers based on NPI and TIN identifiers = 1,351

Claims

Data from the PQRS program were used to perform the correlation analysis for this measure. Data comes from the claims versions of Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation (PQRS #012) and Diabetes: Eye Exam (PQRS #117).

Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation (PQRS #012) demonstrates positive correlation with Diabetes: Eye Exam (PQRS #117)

PQRS #117

Coefficient of correlation = 0.22

Alpha level = 0.05

P-value < 0.001

Number of shared providers based on NPI and TIN identifiers = 11,475

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?*)

Previously Submitted 2013 Data Face Validity Results

The results of the expert panel rating of the validity statement were as follows: N = 16; Mean rating = 4.56 and 87.50% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)

2 – 0 responses

3 – 2 responses (Neither Agree nor Disagree)

4 – 3 responses

5 – 11 responses (Strongly Agree)

Current Testing Data – empirical validity correlation testing results

Registry

Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation has a moderate positive correlation with another evidence-based process of care measure. The correlation is statistically significant at the 95% confidence level and demonstrates the criterion validity of the measure.

Claims

Primary Open-Angle Glaucoma (POAG): Optic Nerve Evaluation has a weak positive correlation with another evidence-based process of care measure. The correlation is statistically significant at the 95% confidence level and demonstrates the criterion validity of the measure. Due to the limited availability of other similar measures reported through claims, we were unable to identify another more suitable candidate for empirical validity.

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*)

Previously Submitted 2013 Registry and Claims Data

With the information available from our data sources, we are unable to determine the type of exception reported, but based on previous studies (see section 2b3.3), levels of exception reporting occur infrequently and are generally valid.

Current Testing Data

Exceptions include:

Documentation of medical reason(s) for not performing an optic nerve head evaluation

Exceptions were analyzed for frequency across providers.

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*)

Previously Submitted 2013 Data

Registry

Of the 1,359 physicians with the minimum (10) number of quality reporting events, there were 233 physicians identified with an exception. There were 4,945 total exceptions reported. The average number of exceptions per physician in this sample is 3.7. The overall exception rate is 3.0%. The exception rate range for the physicians in this sample is 82.9% to 0%.

Claims

Of the 9,616 physicians with the minimum (10) number of quality reporting events, there were 1,249 physicians identified with an exception. There were 16,280 total exceptions reported. The average number of exceptions per physician in this sample is 1.7. The overall exception rate is 1.8%. The exception rate range for the physicians in this sample is 93.9% to 0%.

Current Testing Data

Registry

Amongst the 1,591 providers, there were a total of 2,036 exceptions reported. The average number of exceptions per provider in this sample is 1.0. The proportion of exceptions to patients is 0.01.

Claims

Amongst the 13,685 providers, there were a total of 10,911 exceptions reported. The average number of exceptions per provider in this sample is 1.0. The proportion of exceptions to patients is 0.02.

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*)

Exceptions are necessary to account for those situations when it is not medically appropriate to conduct an optic nerve head evaluation. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for a medical reason. Rather than specifying an exhaustive list of explicit medical reasons for exception for each measure, the measure developer relies on clinicians to link the exception with a specific reason for the decision not to perform the optic nerve evaluation required by the measure.

Some have indicated concerns with exception reporting including the potential for providers to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by providers, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that providers document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each provider's exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that provider. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234.

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.

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of*

p<0.10; correlation of x or higher; patient factors should be present at the start of care) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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*)

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

If stratified, skip to 2b3.9

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

2b3.9. Results of Risk Stratification Analysis:

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

Previously Submitted 2013 Registry and Claims data

Not applicable

Current Testing Data

Not applicable

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

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

Previously Submitted 2013 Registry and Claims data

Measures of central tendency, variability, and dispersion were calculated.

Current Testing Data

Measures of central tendency, variability, and dispersion were calculated.

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*)

Previously Submitted 2013 Data

Registry

Based on the sample of 1,359 included physicians, the mean performance rate is 0.85, the median performance rate is 0.98 and the mode is 1.00. The standard deviation is 0.25. The range of the performance rate is 1.00, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.20 (1.00 – 0.80).

Claims

Based on the sample of 9,616 included physicians, the mean performance rate is 0.77, the median performance rate is 0.86 and the mode is 1.00. The standard deviation is 0.24. The range of the performance rate is 1.00, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.30 (0.95 – 0.65).

Current Testing Data

Registry

Based on the sample of 1,591 included providers, the mean performance rate is 0.94 the median performance rate is 1.0 and the mode is 1.0. The standard deviation is 0.15. The range of the performance rate is 0.95, with a minimum rate of 0.04 and a maximum rate of 1.0. The interquartile range is 0.03 (0.97–1.00). Percentiles are provided below:

Percentile	Value
10th	0.81
25th	0.97
50th	1.00
75th	1.00
90th	1.00

Claims

Based on the sample of 13,377 included providers, the mean performance rate is 0.97 the median performance rate is 1.0 and the mode is 1.0. The standard deviation is 0.11. The range of the performance rate is 0.99, with a minimum rate of 0.01 and a maximum rate of 1.0. The interquartile range is 0.00 (1.00–1.00). Percentiles are provided below:

Percentile	Value
10th	0.96
25th	1.00
50th	1.00
75th	1.00
90th	1.00

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?*)

Previously Submitted 2013 Data

Registry

The range of performance from 0.00 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

Claims

The range of performance from 0.00 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

Current Testing Data

Registry

Outliers are considered to be values less than quartile 1 (0.97) or greater than quartile 3 (1.00) by more than 1.5 the IQR (0.03) and there were approximately 275 outliers in the data set. While only 10% of the data falls below a performance score of 0.81 we believe that there remains meaningful variation across providers' performance considering that scores there include providers that do not pass the measure. Looking at the performance percentiles without outliers, 10% of the data falls below a performance score of 0.97 which demonstrates that there does still exist some variation across providers' performance. See below for performance percentiles with outliers excluded:

Percentile	Value
10th	0.97
25th	1.00
50th	1.00
75th	1.00
90th	1.00

Claims

Outliers are considered to be values less than quartile 1 (1.00) or greater than quartile 3 (1.00) by more than 1.5 the IQR (0.00) and there were approximately 1,667 outliers in the data set. While only 10% of the data falls below a performance score of 0.96 we believe that there remains some meaningful variation across providers' performance considering that scores there include providers that do not pass the measure. Looking at the performance percentiles without outliers, 100% of the data is equal to a performance score of 1.0. See below for performance percentiles with outliers excluded:

Percentile	Value
10th	1.00
25th	1.00
50th	1.00
75th	1.00
90th	1.00

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*)

Previously Submitted 2013 Registry and Claims Data

This test was not performed for this measure

Current Testing Data

This test was not performed for this measure

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*)

Previously Submitted 2013 Registry and Claims Data

This test was not performed for this measure

Current Testing Data

This test was not performed for this measure

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*)

Previously Submitted 2013 Registry and Claims Data

This test was not performed for this measure

Current Testing Data

This test was not performed for this measure

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*)

Previously Submitted 2013 Registry and Claims Data

Data were not available to complete this testing.

Current Testing Data

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

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*)

Previously Submitted 2013 Registry and Claims Data

Data were not available to complete this testing.

Current Testing Data

This test was not performed for this measure. There was no 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*)

Previously Submitted 2013 Registry and Claims Data

Data are not available to complete this testing.

Current Testing Data

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased 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.

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**.

ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

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

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

Attachment:

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.

We have not identified an areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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)*.

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. The AMA, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

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 Merit-based Incentive Payment System (MIPS) https://qpp.cms.gov/mips/quality-measures Merit-based Incentive Payment System (MIPS) https://qpp.cms.gov/mips/quality-measures Payment Program Merit-based Incentive Payment System (MIPS) https://qpp.cms.gov/mips/quality-measures IRIS™ Registry (Intelligent Research in Sight) http://www.aao.org/iris-registry/ MORE Registry (Measures and Outcomes Registry for Eyecare) https://www.aoa.org/more Merit-based Incentive Payment System (MIPS) https://qpp.cms.gov/mips/quality-measures IRIS™ Registry (Intelligent Research in Sight) http://www.aao.org/iris-registry/ MORE Registry (Measures and Outcomes Registry for Eyecare) https://www.aoa.org/more Quality Improvement (Internal to the specific organization) IRIS™ Registry (Intelligent Research in Sight) http://www.aao.org/iris-registry/ MORE Registry (Measures and Outcomes Registry for Eyecare) https://www.aoa.org/more

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

Merit-based Incentive Payment System (MIPS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS). Prior to 2016, this measure was used for Eligible Providers (EPs) in the Physician Quality Reporting System (PQRS). As of 2017, PQRS has been replaced by the Merit-based Incentive Payment System (MIPS). MIPS is a national performance-based payment program that uses performance scores across several categories to determine payment rates for EPs. MIPS takes a comprehensive approach to payment by basing consideration of quality on a set of evidence-based measures that were primarily developed by clinicians, thus encouraging improvement in clinical practice and supporting advances in technology that allow for easy exchange of information.

According to the CY 2019 Quality Payment Program final rule, CMS intends to "make all measures under MIPS quality performance category available for public reporting on Physician Compare in the transition year of the Quality Payment Program, as technically feasible." These measures include those reported via all available submission methods for MIPS-eligible clinicians and groups. This measure has now been included in Physician Compare and Performance Rates will be available in 2019.

The IRIS® Registry (Intelligent Research in Sight) sponsored by the American Academy of Ophthalmology. This is an electronic health record-based comprehensive eye disease and condition registry. It is a centralized data repository and reporting tool that can analyze patient data to produce easy-to-interpret national and inter-practice benchmark reports and provide scientific information to improve public health. The reports can validate the quality of care ophthalmologists provide and pinpoint opportunities for improvement. Eligible physicians who sign up and meet the reporting requirements can use the IRIS Registry to report clinical quality data to the Merit-Based Incentive Payment System. The IRIS Registry will automatically extract and submit data

for MIPS quality measures to the Centers for Medicare & Medicaid Services on behalf of practices integrated with their EHR. Additionally, CMS has confirmed that the IRIS Registry is considered a Clinical Data Registry and a Public Health Registry for the purpose of providing Promoting Interoperability performance points, because of its public health and population health data analyses to improve care.

The Measures and Outcomes Registry for Eyecare (MORE) registry is a qualified clinical data registry (QCDR) sponsored by the American Optometric Association. AOA MORE is the nations' first optometric-focused registry. The primary initial goals of the registry are to assist eye-care practices in improving the quality of care, and to submit quality measures to the Medicare Merit-Based Incentive Payment System (MIPS). As of April 2018, AOA MORE was able to attest and submit data for over 600 optometrists and more than 7,492 AOA members are registered with MORE.

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?*)

The PCPI strongly encourages the use of its measures in quality improvement and accountability initiatives and promotes their use in public reporting programs. Measures developed by the PCPI, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. As a measure developer, we work with measure implementers as opportunities arise to encourage and facilitate the integration of PCPI measures in their programs.

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.*)

Not applicable.

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 PCPI measure development and maintenance process is a rigorous, evidence-based process that has been refined and standardized since the PCPI's inception in 2000. Throughout its tenure, the PCPI has conducted its measure development and maintenance process with strict adherence to several key principles, including the following which underscore the role those being measured have played in the development and maintenance process and in providing feedback based on measure implementation:

Collaborative Approach to Measure Development

PCPI measures are developed and maintained through cross-specialty, multi-disciplinary technical expert panels. Representatives of relevant clinical specialties are invited to participate in our expert panels to advise us throughout the measure development process and as questions arise during measure implementation. Additionally, other health care providers and stakeholders participate in our panels as equal contributors to the measure development process. The PCPI also strives to include on its panels individuals representing the perspectives of patients, consumers, private health plans, and employers. Liaisons from key measure development organizations, including The Joint Commission and NCQA, at times participate in the PCPI's measure development process to ensure measure harmonization. Measure methodologists and coding and informatics experts are also considered important members of the expert panel. This broad-based approach to measure development maximizes the input from those being measured and other stakeholders to develop evidence-based, feasible and clinically meaningful measures.

Public Comment Period

Input from a wide range of stakeholders is integral to the measure development process. To invite other perspectives and expertise beyond the expert panels and particularly from those providers and facilities that will implement these measures, the PCPI submits the measures for public comment. All measures are released for a 30-day public and PCPI member comment period. All comments are reviewed by the technical expert panel to determine whether measure modifications are needed based on comments received.

Feedback Mechanisms

The PCPI has a dedicated mechanism set up to receive measure-related comments and questions from implementers. As comments and questions are received, they are shared with appropriate staff for follow up. If comments or questions require expert input, these are shared with the PCPI's technical expert panels to determine if measure modifications may be warranted. Additionally, for PCPI measures included in federal reporting programs, there is a system that has been set up to elicit timely feedback and responses from PCPI staff in consultation with technical expert panel members, as appropriate.

Feasibility Assessments

The PCPI solicits feedback on measure feasibility in the following domains: data availability, data accuracy, data standards, and workflow to guide future modifications to the measure. During this process, we may receive recommendations to improve the experience of those implementing and reporting on this measure and we follow up on any questions or concerns received by those completing the feasibility assessment. Doing so addresses any issues with interpretation and serves as an important step in the measure development process.

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.

See description in Section 4a2.1.1 above.

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.

As described in Section 4a2.1.1, the PCPI invites feedback through various mechanisms. We obtain input from our topic-specific technical expert panels during the measure development and during the annual maintenance process. Additionally, the PCPI obtains feedback via an online public comment and an email-based process set up to receive measure inquiries from implementers.

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

During the development of this measure, comments were received during the public comment period. An overarching theme from clinicians was a request to clarify what is considered an adequate examination of the optic nerve.

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

Other users expressed concern over the age range of this measure, and that patients younger than 65 may not be covered by Medicare.

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

As a result of clinician feedback to clarify what is considered an adequate examination of the optic nerve, guidance was added to offer further clarity around the intent, while also allowing physicians to use their discretion in selecting the tools to perform optic nerve evaluation.

In response to the concern over age range, the PCPI responded that this measure was developed to align with the clinical practice guidelines, and for implementation and adoption by physicians, payers, and other interested groups to improve quality of care. Although use in the Medicare program may further limit the

patient population to patients over 65, it is possible that other groups would apply the measure to patients over 18 years of age, if appropriate.

Improvement

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

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

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

This measure is intended to promote examination and documentation of the structure and function of the optic nerve, to monitor and detect disease progression among POAG patients. CMS data report a 5 percent decrease in the average performance rate of this measure, from 2013 through 2017. However, reporting rates represent but one facet of the quality improvement process.

While the PCPI creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and/or structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1. Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

4b2. Unintended Consequences

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

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

We have not received reports of unexpected findings resulting from the implementation of this measure. The PCPI has various mechanisms in place for measure users to provide feedback and to identify issues related to the maintenance and implementation of this measure. We convene several topic-specific technical expert panels comprised of various stakeholders including those being measured to advise us regarding any unexpected findings and actions that can be taken to mitigate them.

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

Not applicable.

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)

0563 : Primary Open-Angle Glaucoma: Reduction of Intraocular Pressure by 15% or Documentation of a Plan of Care

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

Not applicable.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

Yes

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

Not applicable.

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 the populations are similar, NQF #0563 measures the reduction in intraocular pressure from the pre-intervention level, while NQF #0086 measures the evaluation of the optic nerve to establish glaucoma disease status and presence of optic nerve damage. This measure intends to monitor, detect, and prevent disease progression among POAG patients. In addition, degeneration of the optic nerve, even while intraocular pressure remains in the normal range, can occur amongst a subtype of open-angle glaucoma patients (normal or low-tension glaucoma). This measure would capture those patients, whereas NQF #0563 would not apply to that patient group.

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): PCPI Foundation

Co.2 Point of Contact: Samantha, Tierney, samantha.tierney@thepcpi.org, 312-224-6071-

Co.3 Measure Developer if different from Measure Steward: PCPI Foundation

Co.4 Point of Contact: Samantha, Tierney, samantha.tierney@thepcpi.org, 312-224-6071-

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.

PCPI measures are developed and maintained under the aegis of topic-specific technical expert panels (TEPs). The PCPI TEPs are comprised of clinicians and other healthcare professionals representing medical specialty societies and other stakeholders. The TEPs provide clinical expertise as well as advise on methodologic questions and review the measures annually to ensure accuracy and adherence to the most current evidence.

Eye Care TEP members include:

John Thompson, MD – TEP Co-Chair

Murray Fingeret, OD

David B. Glasser, MD

Richard Hellman, MD

Mathew W. MacCumber, MD, PhD

Zachary S. McCarty, OD

Parag D. Parekh, MD

Marc Piccolo, OD

Thomas A. Wong, OD

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2006

Ad.3 Month and Year of most recent revision: 04, 2019

Ad.4 What is your frequency for review/update of this measure? Supporting guidelines, specifications, and coding for this measure are reviewed annually

Ad.5 When is the next scheduled review/update for this measure? 04, 2020

Ad.6 Copyright statement:

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

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the PCPI® Foundation (PCPI®) Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for

Performance Improvement® (AMA-PCPI), now known as the PCPI, nor their members shall be responsible for any use of the Measures.

PCPI encourages use of the Measures by other health care professionals, where appropriate.

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

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. The AMA, the PCPI and its members and former members of the AMA-PCPI disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.8 Additional Information/Comments: