NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

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 #: 3617

Corresponding Measures:

Measure Title: Measuring the Value-Functions of Primary Care: Provider Level Continuity of Care Measure

Measure Steward: American Board of Family Medicine

Brief Description of Measure: This is a process measure evaluating primary care physicians; for each physician, their denominator is all of the patients they saw during the evaluation period who had at least 2 PCP visits (could include visits to other PCPs), and the numerator is the number of those patients whose Bice-Boxerman Continuity of Care Index is >= 0.7.

The Bice-Boxerman index is a validated measure of patient-level care continuity that ranges from 0 to 1; 0 reflects completely disjointed care (a different provider for each visit) and 1 reflects complete continuity with the same provider for all visits.

Developer Rationale: This measure seeks to raise the awareness of the importance of continuity of care in primary care. Multiple studies have demonstrated that higher levels of care continuity are associated with lower levels of care utilization and costs. 1-7 By evaluating primary care providers based on the proportion of their patients who exhibit high levels of care continuity, we are signaling to providers that not only is this an important aspect of care, but that it should not fall solely to the patient to ensure their own care continuity. Primary care physicians have a responsibility to assist patients not just during occasional appointments or check-ups, but also across the spectrum of care to promote the highest quality of care for all aspects of health and well-being.

Since Continuity of Care is typically thought of as a characteristic of a patient's experience, this measure first calculates Continuity of Care for each patient using a previously validated index (the Bice-Boxerman index), where patients who have most of their primary care visits to the same provider or a small number of providers have higher Continuity of Care scores (closer to 1.0), while those who see a larger number of different providers have lower Continuity of Care scores (closer to 0.0).

The provider-level measure is the proportion of a provider's seen patients who have a Continuity of Care index of 0.7 or above. The higher this proportion is, the more of the provider's encounters are with patients who see only one (i.e., that provider) or few providers for primary care, thereby indicating that more of the patients seen by that provider experience higher care continuity.

The threshold of 0.7 for the Continuity of Care index was established based on published literature. There are a variety of studies demonstrating that higher levels of care continuity are associated with lower levels of care

utilization and costs for patients. These studies span age groups from pediatric to over age 65, multiple settings (community-dwelling adults, those in long-term care, etc.), different methods for measuring continuity (including the Bice-Boxerman index), and link continuity to multiple types of care utilization (emergency department, hospital).1-7 One study, in particular, that used the Bice-Boxerman index for Primary Care Physician continuity established the 0.7 threshold as associated with statistically significant results.1

Benefits and improvements in quality envisioned:

This measure uses a validated method for calculating a patient's Continuity of Care (the Bice-Boxerman Index); one that has been shown to be linked to outcomes. There are multiple Continuity of Care indices. A previous study of continuity of care in primary care examined 4 different ones and found a strong correlation across them (0.86 to 0.99),1 suggesting that they perform similarly. That publication chose the Bice Boxerman to report their findings because, in part, it appears in another NQF endorsed measure. The Bice Boxerman index has also been used by multiple studies associating care continuity to care utilization and outcomes. 2,4,5 The Continuity of Care (and therefore the overall measure) reflects quality of primary care from the patient perspective, allowing for comparisons of individual clinician's performance to others in their practice or more broadly.

1. Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations. Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

2. Huang ST, Wu SC, Hung YN, Lin IP. Effects of continuity of care on emergency department utilization in children with asthma. Am J Manag Care. 2016 Jan 1;22(1): e31-7.

3. Marshall EG, Clarke B, Burge F, Varatharasan N, Archibald G. Andrew MK. Improving continuity of care reduces emergency department visits by long-term care residents. J Am Board Fam Med. Mar-Apr 2016;29(2):201-8.

4. Kao YH, Tseng TS, Ng YY, Wu SC. Association between continuity of care and emergency department visits and hospitalization in senior adults with asthma-COPD overlap. Health Policy. 2019 Feb; 123(2):222-228.

5. Kao YH, Wu SC. Effect of continuity of care on ED visits in elderly patients with asthma in Taiwan. J Am Board Fam Med. May-Jun 2017;30(3):384-395.

6. Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP. Continuity of care and health care utilization in older adults with dementia in FFS Medicare. JAMA Internal Med. 2016 Sep 1;176(9):1371-8.

7. Ionescu-Ittu R, McCusker J, Ciampi A, Vadeboncoeur AM, Roberge D, Larouche D, Version J, Pineault R. Continuity of primary care and ED utilization among elderly people.

Numerator Statement: The numerator is the number of patients with a continuity index of at least 0.7.

Denominator Statement: The denominator is the total number of patients with continuous enrollment with at least 2 visits to any primary care physicians in the measurement period. The requirement of continuous enrollment ensures that all of the patient encounters will be captured in the data, and the requirement of at least 2 visits is necessary to calculate a Continuity of Care index (the notion of "continuity" isn't applicable to someone who only has 1 physician visit, i.e., there needs to be at least 2 visits to determine if they consistently visit the same or different physicians).

Denominator Exclusions: Since Continuity of Care is about seeing the same clinician, we did not consider patients with only one visit as an exclusion, therefore; we do not have any denominator exclusions.

Measure Type: Process

Data Source: Claims

Level of Analysis: Clinician: Individual

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

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

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

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

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

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

The developer provides the following evidence for this measure:

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

Evidence Summary

- This is a new process measure utilizing claims data at the clinician level to assess the importance of continuity of care in primary care.
- The logic model presented by the developer for this process measure links a physician's knowledge of being evaluated on their patients' continuity of care to an increase in continuity due to ongoing communication, easy-to-use patient portals, email reminders, etc.
- Per developer, the goal of this measure is to raise awareness of the importance of continuity of care in primary care.
- The developer submitted seven studies (USA 2, Canada 2, Taiwan 3) published between 2007 and 2019 demonstrating an association between higher levels of care continuity, including continuity of care with a primary care physician, with lower levels of care utilization and costs.

Question for the Committee:

- How strong is the evidence for this relationship?
- Is the evidence directly applicable to the process of care being measured?

Guidance from the Evidence Algorithm

Process measure based on empirical evidence (Box 3) → Not systematically reviewed or graded (Box 7) → Does the empirical evidence that is summarized include all studies in the body of evidence? YES (Box 8) → High certainty that benefits outweigh undesirable effects (Box 9) → Moderate

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

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

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

- The measure was tested with Optum claims data between 7/1/2018-6/30/2019.
- Clinicians enrolled during the measurement period included 555,213 across all 50 states and Puerto Rico; the number of physicians per state ranged from 393 (AK) to 42,343 (TX) (there are 135 from Puerto Rico).
- The mean performance for the measure was 0.2763 with a standard deviation of 0.3058. The developer notes that a large portion of the providers have a performance outside of the 95% confidence level for mean performance, suggesting statistically significant differences in clinician-level performance.

Disparities

- The developer noted that their data did not allow for an examination into disparities within this cohort but provided some literature addressing disparities in care on the specific focus of measurement.
- The developer cited a 2001 study indicating a close association between disparities in the continuity of care and whether patients were seen in physician offices versus in other settings, like the hospital OP setting or health centers.
- In a different study, the impact of social determinants of health on the ability to access primary care and the delivery of preventive care was explored.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Disparities data was included but literature was provided to support that disparities in care exist. Are you aware of evidence that disparities exist in this area of healthcare?

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

Committee Pre-evaluation Comments:

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

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

- Evidence supports the measure.
- Evidence is moderate based on QQC
- The measure will apply directly to PCP Continuity of Care (COC) and targets PCP population for persons with access to care.
- High level of evidence
- This measure assesses at the provider level the proportion of patients that have a Continuity of Care index of >0.7. In essence, the CCI assesses the number of different providers that a patient sees for primary care. Continuity (seeing a fewer number of unique individuals for primary care) is associated with lower utilization and costs. The measure developers provided 7 studies (2 US, 2 Canada, 3 Taiwan) demonstrating the association b/w higher continuity and lower care utilization and costs.
- The evidence level is moderate
- Evidence is moderate for this measure. The cited studies reference additional factors that impact provider care coordination that are not consistently within the provider's influence or control.
- This process measure quantifies continuity of care using an established formula.
- The measure includes continuity to a “small number” of PCPs, not just 1. What is this “small number,” and was the reliability for this small number tested.
- Many factors other than visiting the same provider are important, including language proficiency, cultural competency, and “looks like me.” Were these factors dealt with in the small number of citations? The citations are largely confined to asthma/COPD and elders, so are people with other conditions included, and if so what is the evidence? Right now I rated the evidence as insufficient but I’m sure there must be more? I am surprised the internal team said evidence was moderate based on these references.
- It is weird to have a measure of primary care at the individual PCP level when this primary care is a team sport. The administration of the office and the team is not just the MD responsibility, to this measure could perpetuate an antiquated care model and assign “blame” to the wrong person.
- I can’t find any evidence that feedback of this measure led to improvement activities as claimed in the logic model – which is not a model but a vague paragraph. Do they know what a logic model is?

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 provided that demonstrates a gap.
- Performance gap exists, that can be improved
- Data is very current and there is a performance gap in the data set they tested. Disparities was not addresses, however they noted evidence in literature regarding SDOH and access to care issues.
- Significant performance gap identified.
- The developer provided data from Optum claims dataset demonstrating the mean performance of 0.2763 (meaning a low amount of continuity) nationally. There is significant spread in scores nationally. Data is not stratified by demographic information to assess performance gap in different populations. The overall data set is 70% Caucasian, 10% Hispanic, 10% African American, and 5% Asian (4% are unknown). The age of included patients varies: 26% are under 35 years old, 21% are aged 34-54 years, 14% aged 55-64 years, 19% aged 65-74 years, and 19% are 75+. The developers provided some references on disparities of care.
- A performance gap exists
- Performance data was provided from seven referenced studies. Each study explored different factors and impacts to care coordination. The mean performance for the measure was 0.2763 with a standard deviation of 0.3058. Results identified many providers that were outside of the 95% confidence level. Gaps and room for improvement were demonstrated. Adjustments were made to allow data from all sites where the practitioner provided care/services. Limited disparity data provided.
- Published studies using the Bice-Boxerman Index show a wide range of values.

- Given the importance of race, ethnicity, literacy, etc., it's unacceptable not to have disparity data on this measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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.

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

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

Evaluators: NQF Staff

Reliability

- Testing was conducted at the performance score level using a beta-binomial model to determine signal to noise.
- Data was collected from the CDM Optum Cliniformatics Data Mart from 7/1/2018 to 6/30/2019. The sample included 555,213 individual providers.
- The mean reliability score was 0.8493.
- The developer reported that as the sample size increased, reliability scores improved, suggesting that the low reliability values may be a function of small sample sizes, as opposed to the inherent reliability of the measure itself.
- The developer states, "While the minimum reliability is 0.2680, after limiting the sample to physicians with >5 and >10 in the denominator, the minimum reliability increased to 0.4878 and 0.6465, respectively, while the mean reliability remained in the mid-0.80s. Small sample sizes can limit the ability to make statistical inferences and increase variability, so that we thought it was reasonable to re-run the analysis after requiring a certain sample size. This suggests that the low reliability values

may be a function of small sample sizes, as opposed to the inherent reliability of the measure itself. Therefore, we interpret these results to indicate that the measure has acceptable levels of reliability."

- The developer states that a reliability score above 0.7 for signal to noise indicates sufficient signal strength to discriminate performance between accountable entities and this measure is very good based on that assumption.

Validity

- Validity testing conducted at the measure score level using empirical validity testing.
- The developer tested the strength of the numerator which defines a threshold of 0.7 for a patient's Continuity of Care index.
    - The developer hypothesized that higher scores on this measure are associated with lower ED visits.
    - The developer reported an odds ratio of 0.72 and concludes that when a provider's continuity of care score of 0.7 or above is associated with decreased ED visits.
- The developer also tested the association of provider performance with patient outcome of at least one ED visit by comparing aggregated ED visits in the provider's denominator with measure performance.
    - The developer hypothesized that physicians with higher measure performance would have a lower percentage of patients with at least one ED visit because they had more continuous care.
    - The developer looked at this association in a model with the provider's measure performance as the independent variable and a model with the provider's measure performance and specialty as the independent variables.
    - In the first model, the developer reported a parameter estimate of -0.11. In the second model, the developer reported a parameter estimate of -0.12. The developer concluded that providers with better scores on the measure have fewer patients with ED visits, even after adjusting for physician specialty.
- The developer also reported that the odds of at least one ED visit was higher for older patients (1.03) and were more common for females (1.318), black Americans (1.335), and those of Hispanic ethnicity (1.029). Patients of Asian descent were significantly less likely to have an ED visit than Caucasians (0.720).

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

Questions for the Committee regarding validity:

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

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

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?

- Data elements are clear. No concerns.
- No major concerns
- Sample size was corrected for therefore a preliminary reliability rating of high. I am interested in the PCP committee members experience with the CDM Optum Cliniformatics Data Mart and Continuity of Care measurements.
- Data elements were clearly defined.
- Mean reliability 0.85. Increased with sample size. Therefore, provider level data for those with small sample sizes may have limited reliability.
- Reliability is high
- Data elements are clearly defined. A large percentage of providers were excluded from the studies. Reliability testing seemed to indicate higher reliability for providers with larger patient volume included in the study. This measure should be able to be consistently implemented.
- Reliability improves as the sample of providers is changed to those having a larger number of patients meeting the numerator criteria. It is assumed that the low reliability values may be a function of small sample sizes, as opposed to the inherent reliability of the measure itself.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure; reliability testing and results for the measure?

- No concerns.
- No major concerns
- Again interested in the input from PCP on their experiences.
- It appears to be reliable but the data is only from one plan
- No. See comments above
- No concerns
- The mean reliability score was 0.8493. The developer reported that as the sample size increased, reliability scores improved. No concerns with reliability.
- No

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

- No concerns
- No major concerns
- no
- No concerns
- no specific concerns about the validity testing. If only 2 visits are required for a patient to be eligible, I would have concerns about attribution at the individual provider level. For example, patient sees provider A for visit 1. Second visit is performed by provider B. Is provider B "penalized" for discontinuity of care even if they have no opportunity to intervene?
- Validity is high
- No concerns with the validity testing conducted by the measure developer.
- Is a 12 month measurement period sufficient to attain a valid measurement of a provider's continuity of care for a population?

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?

- No threats to validity.
- No major concerns
- They used one source of claims data to test and assumed full capture of encounters.
- There are no significant threats to validity. Data elements are clear.
- There is signficant spread in the data indicating meaningful differences. Uses claims data which should capture all data elements.
- No concerns
- Reliable data sources are used for this measure. No concerns about the ability to have complete data for measurement purposes. The measure specifications should result in the ability to identify meaningful differences for quality improvement purposes. Results should be comparable. The measure does not consider patient factors, social determinants of health or racial disparities (70% of patients' data contained race/ethnicity data),
- 2b4 Higher continuity scores parallel low frequencies of ER visits.

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?

- No exclusions.
- No major concerns
- No exclusions and no Risk adjustments
- There is no risk adjustment or case-mix adjustment. This could be helpful in the future.
- There is no risk adjustment. no exclusions
- No concerns
- Measure includes patients with two or more provider visits. Measure steward does not consider patients with one visit to be an exclusion in the measure specifications.
- No risk adjustment

Criterion 3. Feasibility

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

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

- Data elements are generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score).
- All data elements are in defined fields in electronic claims.
- The developer notes that they do not anticipate any difficulties beyond the standard lag time associated with Administrative Claims data.
- The developer notes that there are no fees or other requirements to use any aspect of the measure as specified.

Questions for the Committee:

- 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?

- Data elements are routinely generated during care delivery. No concerns
- No major concerns
- Data elements are available within standard Administrative claims data
- All of the data elements are available and easily collected in the past.
- data elements are generated during routine provision of care. Available in electronic claims. No concerns.
- No concerns
- Data are in the providers' EHR. Data is found in claims data. No concerns with ability to capture data necessary to calculate or report the measure.
- Good

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**

Accountability program details

- The Continuity of Care quality measure has been approved for use in the CMS MIPS program.
- The measure has been used in the PRIME QCDR since the 2018 measurement period. The developer reports that 2020 data is not yet available, although the developer expects similar results to measurement year 2019 (2409 clinicians and 782 TINs used this measure).
- The developer also notes that:

- The measure is being submitted to the CMS MUC list in 2021.
- The Continuity of Care measure will be part of a MIPS Value Pathway (MVP), which is in development.
    - ABFM plans to make the MVP part of its Maintenance of Certification.

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 provides performance results, data and assistance with interpretation via a PRIME Registry Dashboard, written user-guide, annual survey for process improvement, and a PRIME Registry team/Registry vendor team.
- The developer reports that data is captured from EMRs on the cadence defined by the practice, which is typically daily or weekly.
- The PRIME Registry Team sends communications to practices at least quarterly advising them to check their dashboards for accuracy.
- The developer highlights positive feedback from clinicians and no comments regarding burden or unexpected negative consequences related to adoption.
- The developer updated measure specification based on feedback to support clinicians and practices striving to meet patient needs and access to timely care.

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?

☒ **Pass** ☐ **No Pass**

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

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

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

Improvement results

- The developer reports that the measure specifications have been updated and is being submitted as a proportional measure, thus their data from the Registry cannot be used to demonstrate trends at this time.
- The developer will submit the measure to CMS during the self-nomination period opening in July 2021 as a proportional measure and will begin to track trends over time.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer did not identify any unexpected findings.

Potential harms

- The developer did not identify any potential harms.

Questions for the Committee*:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?

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

Preliminary rating for Use:

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?

- Publicly reported and used in accountability programs listed in application
- No concerns
- I need to see the comments from the PCP Committee members.
- These data are routinely available electronic records. I did not see much regarding user feedback
- Has been approved for use in the CMS MIPS program. also used in the PRIME QCDR. Feedback is solicited on the measure/data.
- Publicly reported, currently used in accountability programs
- Measure currently being used in MIPS. How to interpret the results appears to be clear (.7 or higher is an indicator of higher continuity of care with anticipated lower ED visits). Providers were given an opportunity to provide feedback on the measure. Information from the measure steward indicates that this measure is frequently selected for federal performance reporting.
- The Continuity of Care quality measure has been approved for the CMS MIPS program and has been used in the PRIME QCDR since the 2018 measurement period. We obtain feedback with the clinicians/practices through regular one-on-one meetings and through our annual survey.

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.

- No harms
- No major concerns
- I can see that within a closed-loop health system the usability. My concern would be for those practices that are not part of a system or have administrative support.
- It appears that accountability is only determined on the individual physician. No analysis by socioeconomic status, by plan type, by physician groups
- No harms.
- No concerns
- The measure can be used to improve the quality of healthcare. There were no identified or unexpected harms from the measure cited. Using number of provider visits as a proxy for effective care coordination should promote high-quality, efficient healthcare.
- Reporting continuity of care measurements to primary care clinicians seems likely to encourage these clinicians and their patients to value better continuity and that, by some studies (referenced in the application) to improved patient ratings of care and some reduced expenses. No unintended benefits or harms have been recognized.

Criterion 5: Related and Competing Measures

Related or competing measures

- None Identified

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?

- None
- No
- None
- no related or competing measures.
- No related or competing measures
- No competing measures were identified.
- No competing measures

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/10/2021

- No NQF Members have submitted support/non-support choices as of this date.
- No Public or NQF Member comments submitted as of this date.

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3617

Measure Title: Measuring the Value-Functions of Primary Care Provider Level Continuity Measure

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**

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

   Submission document: Testing attachment, section 2a2.2
   - Reliability testing conducted at the measure score level using a signal-to-noise ratio (SNR) analysis.
   - Data was collected from the CDM Optum Cliniformatics Data Mart from 7/1/2018 to 6/30/2019.

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

   Submission document: Testing attachment, section 2a2.3
   - The sample included 555,213 individual clinicians.
   - The developer states that the mean reliability result of 0.8493.
   - For 55,213 providers, the developer reported a minimum score 0.2680 and maximum of 1.
   - Percentiles (10, 25, 50, 75, 90): 0.4938,0.7553, 0.9492, 1, 1.

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 SNR result suggests a high rate of reliability.
    - Box 1: Precise specifications → Box 2: Empirical reliability testing → Box 4: reliability testing conducted on measure score for each measured entity → Box 5: Appropriate testing method → Box 6a. High certainty that measure scores are reliable. → HIGH

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.
    - No concerns. This measure has no denominator exclusions.

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

    Submission document: Testing attachment, section 2b4.
    - The developer calculated the mean, standard deviation, standard error, 95% confidence interval, median, range, and interquartile range of scores for each provider.
    - The developer reports a statistically significant difference in performance among the providers.
    - No concerns.

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.
    - Not applicable.

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

    Submission document: Testing attachment, section 2b6.
    - The developer reported that missing data is not systematic.
    - No concerns.

16. **Risk Adjustment**

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

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

    ☐ Yes ☐ No ☒ Not applicable

    16c. **Social risk adjustment:**

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

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

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

    16d. **Risk adjustment summary:**

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

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

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

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

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

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

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☐ Yes ☐ Somewhat ☐ No (If "Somewhat" or "No", please explain)

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

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

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

☐ Face validity

☒ Empirical validity testing of the measure score

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

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

Submission document: Testing attachment, section 2b2.2

- The developer used logistic regression to test the strength a using threshold of 0.7 for a patient's Continuity of Care index.
- The developer also used logistic regression to test the association of provider performance with a patient outcome of at least one ED visit by comparing aggregated ED visits in the provider's denominator with measure performance. The developer also tested this association by provider specialty.

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

Submission document: Testing attachment, section 2b2.3

- For the threshold test, the developer reported an odds ratio of 0.72 and concluded that when a provider's continuity of care score equal to or above the threshold of 0.7 is associated with decreased ED visits.
- In the first logistic regression, the developer reported a parameter estimate of -0.11. In the second logistic regression, the developer reported a parameter estimate of -0.12. The developer concluded that providers with better scores on the measure have fewer patients with ED visits, even after adjusting for physician specialty.

23. **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)

24. **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)

25. **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.)

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

    Box 1: Potential threats to validity assessed → Box 2: Empirical validity testing conducted using the measure as specified → Box 5: Testing conducted at the measure score level → Box 6: Testing method described and appropriate → Box 7b: High certainty or confidence that the performance measure scores are a valid indicator of quality → HIGH

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **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.**

    - Not applicable

Developer Submission

NQF #: 3617

Corresponding Measures:

De.2. Measure Title: Measuring the Value-Functions of Primary Care: Provider Level Continuity of Care Measure

Co.1.1. Measure Steward: American Board of Family Medicine

De.3. Brief Description of Measure: This is a process measure evaluating primary care physicians; for each physician, their denominator is all of the patients they saw during the evaluation period who had at least 2 PCP visits (could include visits to other PCPs), and the numerator is the number of those patients whose Bice-Boxerman Continuity of Care Index is >= 0.7.

The Bice-Boxerman index is a validated measure of patient-level care continuity that ranges from 0 to 1; 0 reflects completely disjointed care (a different provider for each visit) and 1 reflects complete continuity with the same provider for all visits.

1b.1. Developer Rationale: This measure seeks to raise the awareness of the importance of continuity of care in primary care. Multiple studies have demonstrated that higher levels of care continuity are associated with lower levels of care utilization and costs.1-7 By evaluating primary care providers based on the proportion of their patients who exhibit high levels of care continuity, we are signaling to providers that not only is this an important aspect of care, but that it should not fall solely to the patient to ensure their own care continuity. Primary care physicians have a responsibility to assist patients not just during occasional appointments or check-ups, but also across the spectrum of care to promote the highest quality of care for all aspects of health and well-being.

Since Continuity of Care is typically thought of as a characteristic of a patient's experience, this measure first calculates Continuity of Care for each patient using a previously validated index (the Bice-Boxerman index), where patients who have most of their primary care visits to the same provider or a small number of providers have higher Continuity of Care scores (closer to 1.0), while those who see a larger number of different providers have lower Continuity of Care scores (closer to 0.0).

The provider-level measure is the proportion of a provider's seen patients who have a Continuity of Care index of 0.7 or above. The higher this proportion is, the more of the provider's encounters are with patients who see only one (i.e., that provider) or few providers for primary care, thereby indicating that more of the patients seen by that provider experience higher care continuity.

The threshold of 0.7 for the Continuity of Care index was established based on published literature. There are a variety of studies demonstrating that higher levels of care continuity are associated with lower levels of care utilization and costs for patients. These studies span age groups from pediatric to over age 65, multiple settings (community-dwelling adults, those in long-term care, etc.), different methods for measuring continuity (including the Bice-Boxerman index), and link continuity to multiple types of care utilization (emergency department, hospital).1-7 One study, in particular, that used the Bice-Boxerman index for Primary Care Physician continuity established the 0.7 threshold as associated with statistically significant results.1

Benefits and improvements in quality envisioned:

This measure uses a validated method for calculating a patient's Continuity of Care (the Bice-Boxerman Index); one that has been shown to be linked to outcomes. There are multiple Continuity of Care indices. A previous study of continuity of care in primary care examined 4 different ones and found a strong correlation across them (0.86 to 0.99),1 suggesting that they perform similarly. That publication chose the Bice Boxerman to report their findings because, in part, it appears in another NQF endorsed measure. The Bice Boxerman index has also been used by multiple studies associating care continuity to care utilization and outcomes.2,4,5 The Continuity of Care (and therefore the overall measure) reflects quality of primary care from the patient

perspective, allowing for comparisons of individual clinician's performance to others in their practice or more broadly.

1.Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations. Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

2. Huang ST, Wu SC, Hung YN, Lin IP. Effects of continuity of care on emergency department utilization in children with asthma. Am J Manag Care. 2016 Jan 1;22(1):e31-7.

3.Marshall EG, Clarke B, Burge F, Varatharasan N, Archibald G. Andrew MK. Improving continuity of care reduces emergency department visits by long-term care residents. J Am Board Fam Med. Mar-Apr 2016;29(2):201-8.

4.Kao YH, Tseng TS, Ng YY, Wu SC. Association between continuity of care and emergency department visits and hospitalization in senior adults with asthma-COPD overlap. Health Policy. 2019 Feb;123(2):222-228.

5.Kao YH, Wu SC. Effect of continuity of care on ED visits in elderly patients with asthma in Taiwan. J Am Board Fam Med. May-Jun 2017;30(3):384-395.

6.Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP. Continuity of care and health care utilization in older adults with dementia in FFS Medicare. JAMA Internal Med. 2016 Sep 1;176(9):1371-8.

7.Ionescu-Ittu R, McCusker J, Ciampi A, Vadeboncoeur AM, Roberge D, Larouche D, Version J, Pineault R. Continuity of primary care and ED utilization among elderly people.

S.4. Numerator Statement: The numerator is the number of patients with a continuity index of at least 0.7.

S.6. Denominator Statement: The denominator is the total number of patients with continuous enrollment with at least 2 visits to any primary care physicians in the measurement period. The requirement of continuous enrollment ensures that all of the patient encounters will be captured in the data, and the requirement of at least 2 visits is necessary to calculate a Continuity of Care index (the notion of "continuity" isn't applicable to someone who only has 1 physician visit, i.e., there needs to be at least 2 visits to determine if they consistently visit the same or different physicians).

S.8. Denominator Exclusions: Since Continuity of Care is about seeing the same clinician, we did not consider patients with only one visit as an exclusion, therefore; we do not have any denominator exclusions.

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Clinician : Individual

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

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

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

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

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

NQF_evidence_attachment_Continuity_of_Care_FINAL_4_19_2021.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*)**:**

Measure Title:

Measuring the Value-Functions of Primary Care: Provider Level Continuity of Care Measure

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

Date of Submission: 4/9/2021

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: Continuity of Care

☐ 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.

Evidence has shown that continuity of care (i.e., seeing the same provider repeatedly instead of seeing several different PCPs) is associated with fewer hospitalizations and ED visits. While care continuity is typically thought of as a patient-level measure, our measure is a provider-level measure that quantifies the proportion of their patients who have a "high" level of continuity. Providers who perform better more often have patients who see them as their primary provider.

The logic model is as follows:

A patient has contact with their primary care physician; the physician knows that they will be measured on this patient's continuity of care, so they discuss with the patient the importance of care continuity and encourage them to be back in touch for any of their primary care needs; the practice may also put components into place that encourage continuity through on-going communication, easy-to-use patient portals, email reminders, etc.; patients become more likely to seek out the same provider for future care

and their care continuity increases, and their primary physician has a better understanding of their overall health and well-being, thereby improving patient outcomes.

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

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

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

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review:

- Title
- Author
- Date
- Citation, including page number
- URL

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR. Grade assigned to the **evidence** associated with the recommendation with the definition of the grade. Provide all other grades and definitions from the evidence grading system. Grade assigned to the **recommendation** with definition of the grade. Provide all other grades and definitions from the recommendation grading system.

Body of evidence:

- Quantity – how many studies?
- Quality – what type of studies?

Estimates of benefit and consistency across studies.

What harms were identified?

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?

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.

Continuity of care, typically measured as an index or score based on the number of visits and number of different physicians seen, has been linked to utilization and costs. Studies have shown that higher levels of continuity have been associated with fewer ED utilization in adults and children, and among populations with asthma, COPD, dementia, and those in the long-term care setting (Huang 2016, Marshall 2016, Kao 2017, Kao 2019, Amjad 2016, Ionescu-Ittu 2007, Bazemore 2018).

The study by Ionescu-Ittu et al. specifically examines the impact of continuity to primary care. While set in Canada, the authors report "an increased rate of emergency department use was associated with lack of a primary physician" (defined as low or medium levels of continuity of care with a PCP. Within a US Medicare population with dementia, Amjad et al. found that compared to those in the highest-continuity group, those with the lowest levels of continuity had higher levels of hospitalization, ED visits, CT scans, and overall medical spending. In another study of Medicare patients, Bazemore et al. reported similar results: those in the highest continuity quintile had 14.1% lower healthcare expenditures 16.1% lower hospitalization rates than those in the lowest continuity quintile. Multiple continuity care indices exist. The Bazemore study, which is focused on primary care, associates costs and utilization with 4 different indices, finding them to be highly correlated with one another but ultimately selecting one (the Bice-Boxerman Index) to illustrate results, in part because it is part of another NQF-endorsed measure. Other studies identified also used the same index and found associations to various types of utilizations and outcomes (Huang, 2016; Kao 2017; Kao 2019). Therefore, we felt that the Bice-Boxerman was an appropriate index to use for our measure.

These studies repeatedly demonstrate that high levels of continuity of care, including continuity of care with a primary care physician, are associated with lower care utilization and costs.

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

Primary Care Physicians at ABFM knew the literature and understood the importance of care continuity. They indicated that the Bazemore (2018) article was particularly relevant since it associated 4 different indices with costs and utilization, using the Bice-Boxerman one to illustrate the results in part because it was part of another NQF-endorsed quality measure. To bolster this information, we searched for additional research in PubMed using search terms like "continuity of care" and "continuity of primary care". While not systematic, we explored the search results, eliminating articles whose titles and abstracts indicated that they were clearly not relevant, and then reading through full text versions of others to ascertain their relevance and identify other potential articles (from their references). We were particularly interested in studies utilizing the Bice-Boxerman index, and those specifically focused on primary care. Through this process we identified multiple articles we feel support the use of continuity of care and the Bice-Boxerman index as appropriate for this population.

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

1.Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations. Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

2. Huang ST, Wu SC, Hung YN, Lin IP. Effects of continuity of care on emergency department utilization in children with asthma. Am J Manag Care. 2016 Jan 1;22(1):e31-7.

3.Marshall EG, Clarke B, Burge F, Varatharasan N, Archibald G. Andrew MK. Improving continuity of care reduces emergency department visits by long-term care residents. J Am Board Fam Med. Mar-Apr 2016;29(2):201-8.

4.Kao YH, Tseng TS, Ng YY, Wu SC. Association between continuity of care and emergency department visits and hospitalization in senior adults with asthma-COPD overlap. Health Policy. 2019 Feb;123(2):222-228.

5.Kao YH, Wu SC. Effect of continuity of care on ED visits in elderly patients with asthma in Taiwan. J Am Board Fam Med. May-Jun 2017;30(3):384-395.

6.Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP. Continuity of care and health care utilization in older adults with dementia in FFS Medicare. JAMA Internal Med. 2016 Sep 1;176(9):1371-8.

7.Ionescu-Ittu R, McCusker J, Ciampi A, Vadeboncoeur AM, Roberge D, Larouche D, Version J, Pineault R. Continuity of primary care and ED utilization among elderly people.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

This measure seeks to raise the awareness of the importance of continuity of care in primary care. Multiple studies have demonstrated that higher levels of care continuity are associated with lower levels of care utilization and costs.1-7 By evaluating primary care providers based on the proportion of their patients who exhibit high levels of care continuity, we are signaling to providers that not only is this an important aspect of care, but that it should not fall solely to the patient to ensure their own care continuity. Primary care physicians have a responsibility to assist patients not just during occasional appointments or check-ups, but also across the spectrum of care to promote the highest quality of care for all aspects of health and well-being.

Since Continuity of Care is typically thought of as a characteristic of a patient's experience, this measure first calculates Continuity of Care for each patient using a previously validated index (the Bice-Boxerman index), where patients who have most of their primary care visits to the same provider or a small number of providers have higher Continuity of Care scores (closer to 1.0), while those who see a larger number of different providers have lower Continuity of Care scores (closer to 0.0).

The provider-level measure is the proportion of a provider's seen patients who have a Continuity of Care index of 0.7 or above. The higher this proportion is, the more of the provider's encounters are with patients who see only one (i.e., that provider) or few providers for primary care, thereby indicating that more of the patients seen by that provider experience higher care continuity.

The threshold of 0.7 for the Continuity of Care index was established based on published literature. There are a variety of studies demonstrating that higher levels of care continuity are associated with lower levels of care utilization and costs for patients. These studies span age groups from pediatric to over age 65, multiple settings (community-dwelling adults, those in long-term care, etc.), different methods for measuring continuity

(including the Bice-Boxerman index), and link continuity to multiple types of care utilization (emergency department, hospital).1-7 One study, in particular, that used the Bice-Boxerman index for Primary Care Physician continuity established the 0.7 threshold as associated with statistically significant results.1

Benefits and improvements in quality envisioned:

This measure uses a validated method for calculating a patient's Continuity of Care (the Bice-Boxerman Index); one that has been shown to be linked to outcomes. There are multiple Continuity of Care indices. A previous study of continuity of care in primary care examined 4 different ones and found a strong correlation across them (0.86 to 0.99),1 suggesting that they perform similarly. That publication chose the Bice Boxerman to report their findings because, in part, it appears in another NQF endorsed measure. The Bice Boxerman index has also been used by multiple studies associating care continuity to care utilization and outcomes.2,4,5 The Continuity of Care (and therefore the overall measure) reflects quality of primary care from the patient perspective, allowing for comparisons of individual clinician's performance to others in their practice or more broadly.

1.Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations. Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

2. Huang ST, Wu SC, Hung YN, Lin IP. Effects of continuity of care on emergency department utilization in children with asthma. Am J Manag Care. 2016 Jan 1;22(1):e31-7.

3.Marshall EG, Clarke B, Burge F, Varatharasan N, Archibald G. Andrew MK. Improving continuity of care reduces emergency department visits by long-term care residents. J Am Board Fam Med. Mar-Apr 2016;29(2):201-8.

4.Kao YH, Tseng TS, Ng YY, Wu SC. Association between continuity of care and emergency department visits and hospitalization in senior adults with asthma-COPD overlap. Health Policy. 2019 Feb;123(2):222-228.

5.Kao YH, Wu SC. Effect of continuity of care on ED visits in elderly patients with asthma in Taiwan. J Am Board Fam Med. May-Jun 2017;30(3):384-395.

6.Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP. Continuity of care and health care utilization in older adults with dementia in FFS Medicare. JAMA Internal Med. 2016 Sep 1;176(9):1371-8.

7.Ionescu-Ittu R, McCusker J, Ciampi A, Vadeboncoeur AM, Roberge D, Larouche D, Version J, Pineault R. Continuity of primary care and ED utilization among elderly people.

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

The CDM Optum Clinformatics Data Mart (CDM, SES version 3.0) was used for testing of this measure. We used the most recent 12 months (7/1/2018-6/30/2019) of Optum claims. The Optum CDM is a database comprised of administrative claims for members of a larger national managed care company affiliated with Optum. The CDM contains medical claims and lab results for 15 to 18 million annual covered lives spanning 50 states. These data have been reviewed by external statisticians and found to be in compliance with the HIPPA Privacy Rules and considered de-identified. Optum restricts access and use of the CDM to clients with agreement. The ABFM has signed data use agreement with Optum which grants ABFM access to the CDM via the Stanford PHS, whose Data Core team reviews and approves access for individual user. For those included and enrolled during the measurement period, all primary care physician encounters should be captured and included in these data.

The measured entity is the individual clinician. The total number of clinicians included is 555,213. Geographically, all 50 states (and Puerto Rico) are represented, with the number of physicians per state ranging from 393 (AK) to 42,343 (TX) (there are 135 from Puerto Rico).

A total of 5,478,835 patients are included in the analysis. They are more often female (58%), and their race/ethnicity breakdown is: 70% Caucasian, 10% Hispanic, 10% African American, and 5% Asian (4% are unknown). The age of included patients varies: 26% are under 35 years old, 21% are aged 34-54 years, 14% aged 55-64 years, 19% aged 65-74 years, and 19% are 75+.

Descriptive statistics for the performance measure scores for the clinician (the measured entity) are as follows:

N=555213

Mean=0.2763

Std Dev=0.3058

Lower 95% CL for Mean=0.2755

Upper 95% CL for Mean=0.2771

Lower Quartile=0

Upper Quartile=0.5000

Median=0.1802

Min=0

Max=1

Performance Measure Score by Decile:

10th Pctl=0

20th Pctl=0

30th Pctl=0

40th Pctl=0.0769

50th Pctl=0.1801

60th Pctl=0.2857

70th Pctl=0.4042

80th Pctl=0.5238

90th Pctl=0.7500

Results are interpreted as showing a significant spread between the minimum and maximum scores (0 to 1), as well as the median and minimum (0 to 0.18), median and maximum (0.18 to 1), and the interquartile range (.00 to .50). Additionally, as evidenced by the percentiles, a large portion of the providers have a performance outside of the 95% confidence level for mean performance, suggesting statistically significant differences in clinician-level performance.

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.

Not applicable.

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.*

Our data did not allow for an examination into disparities within this cohort. However, there is some literature that may be relevant discussed in 1b.5 below.

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

Health disparities regarding continuity of care have received limited study. However, in 2001 a study was published where almost 35,000 US adults were surveyed about their continuity of care. The results revealed disparities in the identification of a regular site of care and continuity of care with the same provider within a site. However, disparities in continuity of care were closely associated with whether patients were seen in physician offices versus in other settings, like the hospital OP setting or health centers.1 Disparities in access to care, including primary care, have previously been demonstrated. Social determinants can negatively impact the ability to access primary care and the delivery of preventive care.2 It is reasonable to infer that disparities in access to and receipt of primary care services would also negatively impact continuity of care, suggesting that improving continuity of care requires strategies to increase access to vulnerable populations."

1. Doesher MP, Saver BG, Fiscella K, Franks P. Racial/ethnic inequities in continuity and site of care: location, location, location. Health Serv Res. 2001 Dec;36(6 Pt 2):78-89.

2. Katz A, Chateau D, Enns JE, Valdivia J, Taylor C, Walld R, McCulloch S. Association of the social determinants of health with quality of primary care. Ann Fam Med. 2018 May;16(3):217-224

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

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

We currently do not have a measure specific web page.

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:** Data_Dictionary_Continuity_of_Care.xlsx

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

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

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

Not an instrument-based measure

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

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable.

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

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

The numerator is the number of patients with a continuity index of at least 0.7.

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

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

The numerator equals the number of eligible patients who have a Bice-Boxerman continuity index score of at least 0.7 during the measurement time period.

For each patient, the continuity index score is calculated using the Bice-Boxerman Continuity of Care calculated as follows: Bice Boxerman-Continuity of Care Patient = See Appendix A.1 page 2 for calculation (since the NQF system only allows HTML text, it strips “special characters” in formulas, thus we had to add the calculation to the appendix).

Where k is the number of providers, n_i is the number of visits to provider i, and N is the total number of visits. (Note that it is necessary that the patient has at least two visits.)

The index can range from 0 to 1, the higher the number the greater the Continuity of Care. If someone has all of their visits with a single provider, their index would equal 1; while someone who saw a different provider for each visit (e.g., 1 visit each to 2 or more providers) would have an index of 0. Someone who saw one provider 5 times and a second provider 1 time would have an index equal to 0.67.

Compared to lower scores (e.g., 0.6 or lower), continuity index scores of 0.7 or higher have been associated significantly lower Medicare expenditures and significantly lower odds of hospitalization1.

1.Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations.

Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

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

The denominator is the total number of patients with continuous enrollment with at least 2 visits to any primary care physicians in the measurement period. The requirement of continuous enrollment ensures that all of the patient encounters will be captured in the data, and the requirement of at least 2 visits is necessary to calculate a Continuity of Care index (the notion of “continuity” isn’t applicable to someone who only has 1 physician visit, i.e., there needs to be at least 2 visits to determine if they consistently visit the same or different physicians).

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets –*

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)

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

For each physician, the denominator is calculated by summing the total number of patients with two or more primary care visits who had at least one of those visits with that physician. This means if a patient saw more than one PCP, they would be in the denominator for each of those PCPs. When using claims, patients must have continuous enrollment over the measurement period, i.e., from 2018-07-01 to 2019-06-30.

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

Since Continuity of Care is about seeing the same clinician, we did not consider patients with only one visit as an exclusion, therefore; we do not have any denominator exclusions.

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.)*

Not applicable

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.)*

No stratification of measure results is required.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

Step 1: Identify all patients with at least 2 visits to a Primary Care Provider in either the office or outpatient setting. In the Optum data, this reflects the situation where a claim indicates that a primary care physician was seen and the place of service is in office or other outpatient place of service. This is done using the health care services categorization code (i.e., HCCC=01) to identify primary care physicians, and the place of service codes

(i.e., POS= 01,02,03,04,11,12,13,14,15,16,17,41,42,49,50,53,57,60, or 71). More detail is provided in the data dictionary.

Step 2: Retain the unique physician identifier (NPI) associated with each visit for the patients in step 1. A patient will appear in the denominator for each physician they see during the time period (i.e., if someone

sees Dr. “A” once and Dr. “B” three times, that patient will appear in the denominator for Dr. A and the denominator for Dr. B).

Step3: Calculate patient continuity index score using the Bice-Boxerman calculation as follows:

Bice-Boxerman-Continuity of Care Patient = See Appendix A.1 page 2 for calculation (since the NQF system only allows HTML text, it strips “special characters” in formulas, thus we had to add the calculation to the appendix).

Where k is the number of providers, n_i is the number of visits to provider i, and N is the total number of visits. Note that it is necessary that the patient has at least two visits.

So, in the example above, the patient who saw Dr. A once and Dr. B three times would have a Bice-Boxerman Continuity of Care index of: [(12 + 3^2)] – 4 / 12 = 0.5. Some simple calculations would show that if this person had only seen Dr. B for all 4 visits their Continuity of Care index would be = 1.0, and similarly, if another visit was added to another PCP (Dr. C), their Continuity of Care index would be less than 0.5, reflecting their experience of more disparate care.

Step 4: Determine if the patient level continuity has Met or Not Met the 0.7 threshold. For each patient, if their index is >=0.7 then they are included in the numerator. In the above example, the patient (using the original scenario) would be in the denominator for both Dr. A and Dr. B, but would NOT be in either numerator.

Step 5: Divide the numerator by the denominator. This reflects the proportion of patients that provider saw who have a Continuity of Care index of at least 0.7.

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.

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 a survey/patient reported data.

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

If other, please describe in S.18.

Claims

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.

Administrative claims data.

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 : Individual

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

Outpatient Services

If other:

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

Not applicable

2. Validity – See attached Measure Testing Submission Form

NQF_testing_attachment_FINAL_4_19_2021.docx

2.1 For maintenance of endorsement

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

2.2 For maintenance of endorsement

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

2.3 For maintenance of endorsement

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

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

Measure Number (*if previously endorsed*)**:**
Measure Title: **Measuring the Value-Functions of Primary Care: Provider Level Continuity of Care Measure**
Date of Submission: **4/9/2021**

Type of Measure:

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

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*).

The CDM Optum Clinformatics Data Mart (CDM, SES version 3.0) was used for testing of this measure. We used the most recent 12 months (7/1/2018-6/30/2019) of Optum claims. The Optum CDM is a database comprised of administrative claims for members of a larger national managed care company affiliated with Optum. The CDM contains medical claims and lab results for 15 to 18 million annual covered lives spanning 50 states.

Compared to the US Census 2010, patients in the Optum Clinformatic Data Mart (CDM) are representative of the gender and age distribution of US population. The proportion of patients is identical to the Census in Midwest and West, but lower in Northeast (10% vs. 18% Census) and higher in South (43% vs. 37% Census). The proportional comparison of patient ethnicity to Census is challenging given that the ethnicity data exist for only 70% of patients in CDM. Nevertheless, the rankings of White, Hispanic, African American and Asian are the same as they are in the Census. The codebook and address or zip code level information is not available for either patient or provider in this dataset, thus we do not have a breakdown of rural versus urban patient or provider characteristics, however; based on all other factors being consistent with US Census data, we don't expect a rural / urban disproportion in the Optum data set.

The codebook and address or zip code level information is not available for either patient or provider in this dataset, thus we do not have a breakdown of rural versus urban patient or provider characteristics.

These data have been reviewed by external statisticians and found to be in compliance with the HIPPA Privacy Rules and considered de-identified. Optum restricts access and use of the CDM to clients with agreement. The ABFM has signed data use agreement with Optum which grants ABFM access to the CDM via the Stanford PHS, whose Data Core team reviews and approves access for individual user. For those included and enrolled during the measurement period, all primary care physician encounters should be captured and included in these data.

1.3. What are the dates of the data used in testing? **7/1/2018-6/30/2019**

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*)

The measured entity is the individual physician. The total number of physicians included is 555,213. Geographically, all 50 states (and Puerto Rico) are represented, with the number of physicians per state ranging from 393 (AK) to 42,343 (TX) (there are 135 from Puerto Rico).

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*)

A total of 5,478,835 patients are included in the analysis. They are more often female (58%), and their race/ethnicity breakdown is: 70% Caucasian, 10% Hispanic, 10% African American, and 5% Asian (4% are unknown). The age of included patients varies: 26% are under 35 years old, 21% are aged 34-54 years, 14% aged 55-64 years, 19% aged 65-74 years, and 19% are 75+.

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

The same data were used for all aspects of testing.

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. Sociodemographic information such as income, education, and language were not assessed.

The Optum claims database does not include social risk factors such as income, education, language, or community characteristics. Therefore, social risk factors were not available for the 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*)

To assess signal-to-noise, we employed the beta-binomial model as described by JL Adams in "The Reliability of Provider Profiling" (1). For each PCP we had numerators and denominators. Through the estimation of the beta-binomial parameters (often referred to as alpha and beta) as described by Adams (1), we estimated the provider-to-provider variance and the within-provider variance (simply the binomial variance for each provider). The ratio of these estimates then produced an estimate of the reliability at each facility, where a reliability of 0 implies that all variability is due to measurement error, while a reliability of 1 indicates that all variability is due to real differences in performance. The distribution of reliability estimates across all facilities was examined.

Because this measure is only applicable for patients who have more than one PCP visit during the measurement period, this can limit the number of patients eligible for the denominator for some PCPs (because any of their patients with less than 2 visits are not included), and therefore many PCPs had small sample sizes (denominators < 5). Therefore, in addition to examining the reliability for the entire population, we also examined the reliability for the subsets of providers who had denominators of greater than 5 and greater than 10.

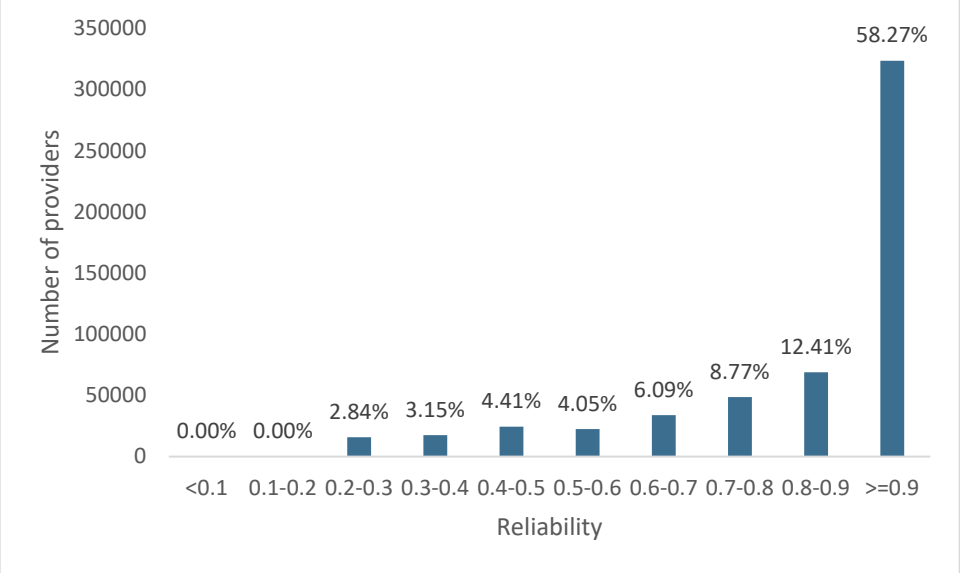
1. Adams, JL. The reliability of provider profiling: A tutorial. RAND Health, 2009.

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

The reliability testing for the entire population of eligible providers produced a mean reliability of 0.8493 and a median reliability of 0.9492. Other percentiles are presented here:

N	Minimum	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	Maximum	Mean
555213	0.2680	0.4938	0.7553	0.9492	1	1	1	0.8493

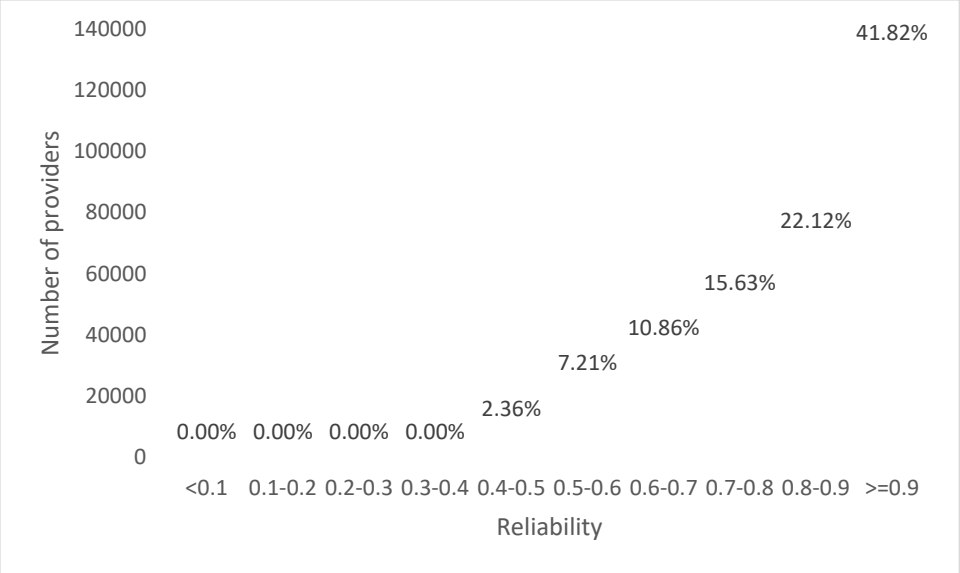
A graphical representation of the distribution is as follows:



We noted that the minimum reliability was only 0.2680, and we wanted to understand whether that was a function of the measure or a function of the sample size for some PCPs. When we re-ran the reliability after limiting to physicians with denominators >= 5 we observed:

N	Minimum	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	Maximum	Mean
311580	0.4878	0.6095	0.7233	0.8672	0.9556	1	1	0.8323

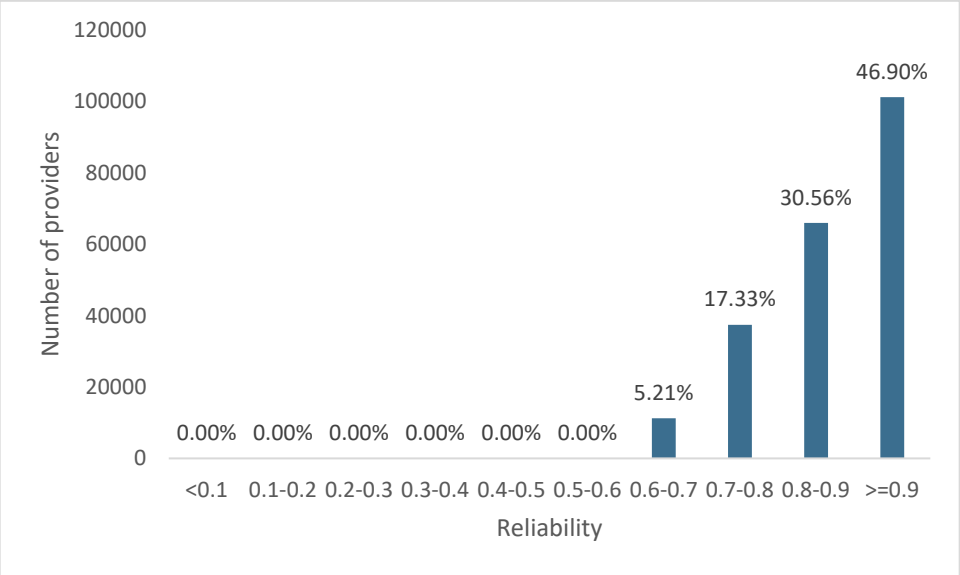
A graphical representation of the distribution is as follows:



And when we limited it to physicians with denominators >= 10 we observed:

N	Minimum	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	Maximum	Mean
215839	0.64648	0.73615	0.81201	0.89284	0.95062	0.98589	1	0.87476

A graphical representation of the distribution is as follows:



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?*)

As previously noted, a reliability of 0.7 is generally viewed as an acceptable threshold. Our mean overall reliability of 0.8493 is very good, suggesting that the measure is highly reliable overall. While the minimum reliability is only 0.2680, when we re-ran the reliability after limiting the sample to physicians with >5 and >10 in the denominator, the minimum reliability increased to 0.4878 and 0.6465, respectively, while the mean reliability remained in the mid-0.80s. Small sample sizes can limit the ability to make statistical inferences and increase variability, so that we thought it was reasonable to re-run the analysis after requiring a certain sample size.

This suggests that the low reliability values may be a function of small sample sizes, as opposed to the inherent reliability of the measure itself. Therefore, we interpret these results to indicate that the measure has acceptable levels of reliability.

It should be noted that the reliability analyses limiting the sample to those providers with >=5 and >=10 patients excluded 44% and 61%, respectively, of the providers. However, the motivation for including these analyses was to explore how reliability changed with sample size. Changes in the distribution percentiles likely reflect fewer extreme values (both 0.0s and 1.0s) that are common among small sample sizes. That is: in addition to the minimum and 10th percentiles increasing as sample size was limited, the 50th, 75th, and 90th percentiles were reduced, likely because of fewer values of 1.0 as sample size was reduced. However, these shifts were relatively small given the large proportion of providers who were excluded, suggesting that the middle and upper part of the distribution were similar across the samples analyzed. Additionally, the mean reliability increased as the sample size was restricted.

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)*
We tested two levels of validity for this measure. First, we explored the validity of using 0.7 as a threshold for the patient-level Continuity of Care within the Optum database used for these analyses. Second, we explored the validity of provider-level performance by examining its association to patient outcomes.

To test the validity of the threshold of 0.7, we used logistic regression to associate the odds of at least one ED visit in the measurement year with achieving the threshold of 0.7. ED visits were chosen as an outcome of interest based on the extensive literature that has previously associated higher levels of continuity with lower levels of utilization, including ED utilization.1-7 We adjusted the logistic model for gender and race. This is designed to look for evidence of convergent validity: the higher Continuity of Care measure scores associated with a lower odds of an ED visit.

Using the same logic, we examined the association of aggregated ED visits among those in each provider's denominator with that provider's performance on this measure. We did this by calculating, for each provider, the percentage of patients in their denominator who had at least one ED visit during the measurement period. Our hypothesis was that physicians with higher measure performance (i.e., a greater percentage of patients with more continuous care) would have a lower percentage of patients with at least one ED visit.

To examine this association, we used linear regression (PROC REG in SAS) to estimate two models, each using the percent with at least 1 ED visit as the response variable. The first model used provider measure performance as the only independent variable; the second model also included physician specialty as an independent variable, to account for differences across specialties. For the second model, we used reference coding (with Pediatrics as the reference). Given that the "Other" category included more than one specialty (like "Maternal Specialist" and "Adult Medical Specialist") but accounted for only 7% of the providers, it these providers were excluded from this analysis.

1.Higher Primary Care Physician Continuity is Associated with Lower Costs and Hospitalizations. Bazemore et al. Annals of Family Medicine. 2018. 16, 492-497.

2. Huang ST, Wu SC, Hung YN, Lin IP. Effects of continuity of care on emergency department utilization in children with asthma. Am J Manag Care. 2016 Jan 1;22(1):e31-7.

3.Marshall EG, Clarke B, Burge F, Varatharasan N, Archibald G. Andrew MK. Improving continuity of care reduces emergency department visits by long-term care residents. J Am Board Fam Med. Mar-Apr 2016;29(2):201-8.

4.Kao YH, Tseng TS, Ng YY, Wu SC. Association between continuity of care and emergency department visits and hospitalization in senior adults with asthma-COPD overlap. Health Policy. 2019 Feb;123(2):222-228.

5.Kao YH, Wu SC. Effect of continuity of care on ED visits in elderly patients with asthma in Taiwan. J Am Board Fam Med. May-Jun 2017;30(3):384-395.

6.Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP. Continuity of care and health care utilization in older adults with dementia in FFS Medicare. JAMA Internal Med. 2016 Sep 1;176(9):1371-8.

7.Ionescu-Ittu R, McCusker J, Ciampi A, Vadeboncoeur AM, Roberge D, Larouche D, Version J, Pineault R. Continuity of primary care and ED utilization among elderly people.

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

The results of the validation of the threshold of 0.7 for the patient-level Continuity of Care indicated that achieving that threshold was significantly associated with a decreased odds of having one or more ED visit (adjusted Odds Ratio = 0.72, p<.0001). The table below shows the odds ratio (OR) and 95% confidence limits of an ED visit for those with a Continuity of Care index of 0.7 or higher. The results also suggest that the odds of at least one ED visit increased with age and were more common for females (vs males) and those of Black race (vs Caucasians) and Hispanic ethnicity. Those of Asian decent were significantly less likely to have an ED visit than Caucasians.

Odds Ratio Estimates from Adjusted Logistic Regression (response variable: 1 or more ED visits during the measurement period)

Effect	Reference	Estimate	95% Wald Confidence Interval
CONTINUITY OF CARE > 0.7	CONTINUITY OF CARE < 0.7	0.718	(0.715, 0.721)
Age (years)	Lowest Age	1.03	(1.03, 1.03)
Female Gender	Male	1.318	(1.312, 1.324)
Unknown Gender	Male	1.317	(0.826, 2.099)
Black Race	Caucasian	1.335	(1.326, 1.344)
Asian Race	Caucasian	0.720	(0.720, 0.729)
Unknown Race	Caucasian	0.972	(0.961, 0.983)
Hispanic Ethnicity	Non-Hispanic	1.029	(1.021, 1.037)

The results of the measure-level validity are as follows. Below is the mean (SD) percentage (the response variable) by the only provider-specific characteristic available in the data: specialty type.

Specialty	Number of providers	Mean (SD) Percentage of Patients in Measure Denominator with 1+ ED visit
Family Practice	255,257	31.6% (31.2%)
General IM	166,526	31.9% (30.9%)
OB	61,097	28.8% (29.9%)
Pediatrics	31,636	3.9% (8.9%)
Other	40,697	22.2% (29.1%)

The results of the first model that used provider measure performance as the only independent variable are below:

ANOVA Table

Source	DF	SS	MS	F	p-value
Model	1	674.98043	674.98043	7244.13	<0.0001
Error	555211	51732	0.09318		
Corrected Total	555212	52407			

Parameter Estimates

Variable	DF	Estimate	SE	t	p-value
Intercept	1	0.323	0.0006	584.26	<0.0001
Measure Performance	1	-0.114	0.0013	-85.11	<0.0001

The results of the second model, which also included physician specialty as an independent variable, are below:

ANOVA Table Continues

Source	DF	SS	MS	F	p-value
Model	4	2958.487	739.62	8310.17	<0.0001
Error	514511	45792.49	0.089		
Corrected Total	514515	48750.98	0	0	0
0	DF	Type III SS	MS	F	p-value
Measure Performance	1	678.07	678.97	7618.64	<0.0001
Physician Specialty	3	2291.68	763.90	8582.91	<0.0001

Parameter Estimates

Variable	DF	Estimate	SE	t	p-value
Intercept	1	0.078	0.0017	45.01	<0.0001
Measure Performance	1	-0.120	0.0014	-87.28	<0.0001
Family Practice	1	0.272	0.0018	152.64	<0.0001
General IM	1	0.281	0.0018	153.55	<0.0001
OB	1	0.225	0.0021	1.07.72	<0.0001

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

These results indicate that those with a Continuity of Care >= 0.7 have significantly lower odds of having an ED visit in the measure year, which demonstrates convergent validity of the Continuity of Care with this

validation metric. One can infer from this that higher PCP performance measure scores (which reflect a higher percentage of patients with a Continuity of Care >0.7) is associated with fewer ED visits on average for their patients.

In both models, Measure Performance is statistically significant, with a negative parameter estimate (-0.11 in the first model and -0.12 in the second model). This indicates that higher provider performance on the quality measure is significantly negatively associated with the percent of patients in their denominator who had an ED visit. That means that providers who score higher (better) on the measure have fewer patients with ED visits, indicating that better measure performance is associated with better patient outcomes, even after adjusting for physician specialty.

_____

2b2. EXCLUSIONS ANALYSIS

☒ **no exclusions — *skip to section 2b4***

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

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*)

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*)

_____

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

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

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

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

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

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

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

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

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

Descriptive statistics for the performance measure scores for all tested entities (physicians) were constructed. These statistics include the mean, standard deviation and standard error, 95% confidence interval, median, range, and interquartile range of scores across the measured entities.

Meaningful difference is defined as a significant spread (>20%) between minimum and maximum scores or a significant spread between median and minimum scores, median and maximum scores, and/or the interquartile range.

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*)

Descriptive statistics for the performance measure scores are as follows:

N	Mean	Std Dev	Lower 95% CL for mean	Upper 95% CL for mean	Lower Quartile	Upper Quartile	Median	Min	Max
555213	0.2763	0.3058	0.2755	0.2771	0	0.5000	0.1802	0	1

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?*)

Results are interpreted as showing a significant spread between the minimum and maximum scores (0 to 1), as well as the median and minimum (0 to 0.18), median and maximum (0.18 to 1), and the interquartile range (.00 to .50). Additionally, as evidenced by the percentiles, a large portion of the providers have a performance outside of the 95% confidence level for mean performance, suggesting statistically significant differences in provider-level performance.

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

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

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

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

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

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

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

We used claims data, which should capture all encounters with PCPs for those included in the database, so there shouldn't be any missing data; if a covered individual had a PCP visit it should appear in the claims.

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*)

We used claims data, which should capture all encounters with PCPs for those included in the database, so there shouldn't be any missing data; if a covered individual had a PCP visit it should appear in the claims.

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*)

We used claims data, which should capture all encounters with PCPs for those included in the database, so there shouldn't be any missing data; if a covered individual had a PCP visit it should appear in the claims.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

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

Not Applicable.

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 do not anticipate any difficulties beyond the standard lag time associated with Administrative Claims data.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or other requirements to use any aspect of the measure as specified.

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)

Payment Program
CMS Quality Payment Program
https://qpp.cms.gov/mips/explore-measures?tab=qualityMeasures&py=2020#measures

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

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

The Continuity of Care quality measure has been approved for the CMS MIPS program and has been used in the PRIME QCDR since the 2018 measurement period. For measurement year 2019 we had 2409 clinicians and 782 TINs use this measure. We have not received 2020 data, however; expects similar results.
The Continuity of Care is being submitted to the CMS MUC list in 2021.
We are also developing a MIPS Value Pathway (MVP), of which the Continuity of Care measure will be a part. We are currently collaborating with CMS to finalize our MVP for measurement year 2022. The ABFM plans to make the MVP part of our Maintenance of Certification. While we do not have a "go live" date yet, the MVP is on our roadmap.
Name of program and sponsor: CMS Quality Payment Program Merit Based Incentive Payment System
Purpose: "To improve the care received by Medicare beneficiaries. To lower costs to the Medicare program through improvement of care and health. To advance the use of healthcare information between allied providers and patients. To educate, engage and empower patients as members of their care team."
Geographic area and number and percentage of accountable entities and patients included: This is a national program for CMS applying to all clinicians and practices who receive CMS payments.
Level of measurement and setting: Individual/Clinician – group/practice

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable.

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.

Performance results, data and assistance with interpretation are provided to those being measured (or other users) in the following ways:

1. PRIME Registry Dashboard: allows the clinician to view measure performance at the practice, clinician or location level. The measure is listed with the registry ID number and the option for the clinician to mark this

measure as a favorite (Since the NQF system only allows HTML text, we had to add our screenshots with descriptions in Appendix A.1 pages 2-7).

There is a hover feature that will provide additional information about the measure:

- MIPS eligibility
- Measure definition
- Link to the full measure specification
- Arrow up indicating higher score is better
- Link for the clinician to raise a service ticket if they have questions/concerns about their measure.
- There is an export feature allowing clinicians to download the data in pdf, csv, or excel format.
- There is a link to provide greater detail on the measure such as numerator, denominator, etc.
- The dashboard also provides the clinician with the measure score timeframe and the date the data was last refreshed.
- Additionally, the PRIME Registry provides the practice with a user-friendly "User Guide" and links to helpful resources

2. Written user-guide: available in the dashboard, provides guidance on how to navigate and interpret data contained within the dashboard (see Appendix A.1 pages 2-7 for screenshots).

3. The PRIME Registry team/Registry vendor team: works with clinicians and practice administrators on interpreting measure specifications and reviewing measure results and benchmarks. Content experts from mapping, coding, calculation, and specification teams are convened to address questions and provide education. Each practice reviews their specific measures with the Registry vendor. Every clinician/practice/location in PRIME receives Continuity of Care scores.

4. Annual survey: to identify potential opportunities for process improvement, each practice receives a survey on patient experience. The feedback and data from the survey is shared with the Registry team and Registry vendor to address improvements.

We did not use a sample of measured entities. Clinicians and Practices have been using the Continuity of Care quality measure in the PRIME QCDR registry since 2018. In 2020, 2,409 clinicians and 782 TINS in the PRIME Registry - representing patients across 50 states and practices in 47 states - used the Continuity of Care measure for quality reporting.

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.

Data is captured from EMRs on the cadence defined by the practice, which is typically daily or weekly. Once a month, all that data is refreshed in the user dashboard. The PRIME Registry Team sends communications to practices at least quarterly advising them to check their dashboards for accuracy. Our Registry team and vendor personnel work with the practices on their measure results/interpretation and, if needed, correct any issues.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

We have not had any clinicians or practice administrators report measure use, implementation, or interpretation burden. While implementing and testing this measure, clinicians found the measure results meaningful which led to over 2,400 clinicians and 782 TINs to choose the measure to report to CMS as part of the MIPS program in 2020.

We obtain feedback with the clinicians/practices through regular one-on-one meetings and through our annual survey.

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

Positive feedback was received from clinicians that the measure is meaningful to themselves and their patients. We did not receive any comments regarding burden or unexpected negative consequences related to adoption.

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

No negative feedback was received from other users.

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.

Early in development, the measure specifications excluded clinicians practicing in multiple locations. Practices with multiple locations shared concern that continuity was not being captured for patients who see the same clinician but at different practice locations. Developers updated measure specification to support clinicians and practices striving to meet patient needs and access to timely care.

Improvement

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

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

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

The PRIME Registry contains three years of data, however; the Continuity of Care measure in PRIME is currently specified as a continuous variable. We have updated the measure specifications and are submitting this measure as a proportional measure, thus our data from the Registry cannot be used to demonstrate trends at this time. We will submit the Continuity of Care measure to CMS during the self-nomination period opening in July 2021 as a proportional measure and will begin to track trends over time.

Although not required for initial endorsement, we will follow the performance results to further the goal of high-quality, efficient healthcare for individuals and populations. Primary Care clinicians value Continuity of Care and are more likely to engage with their patients on more than just the chief complaint for which the patient originally sought care. This encourages patients to see their Primary Care clinician regularly and for all concerns. Additionally, primary care clinicians often times have intrinsic motivation to improve patient outcomes through Continuity of Care which aligns with CMS Improvement Activity: IA_PM_12 - Improvement Activity Title: Population Empanelment addresses Continuity of Care.

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.

No unexpected findings to date.

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

No unexpected benefits to date.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

No

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):
Are the measure specifications harmonized to the extent possible?
5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable.

Appendix

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

Attachment **Attachment:** Appendix_A.1_FINAL_4_19_2021.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American Board of Family Medicine

Co.2 Point of Contact: Jill, Shuemaker, jshuemaker@theabfm.org, 202-600-9447-

Co.3 Measure Developer if different from Measure Steward: American Board of Family Medicine

Co.4 Point of Contact: Jill, Shuemaker, jshuemaker@theabfm.org, 202-600-9447-

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.

Andrew Bazemore, MD, MPH, American Board of Family Medicine, Lexington, Kentucky: Conceptualization

Stephen Petterson, PhD, Robert Graham Center for Policy Studies, Washington, DC: Conceptualization

Lars E. Peterson, MD, PhD, American Board of Family Medicine, Lexington, Kentucky

Department of Family and Community Medicine, University of Kentucky, Lexington, Kentucky: Conceptualization

Yoonkyung Chung, PhD, Robert Graham Center for Policy Studies, Washington, DC: Conceptualization

Robert L. Phillips Jr, MD, MSPH American Board of Family Medicine, Lexington, Kentucky: Conceptualization

Ming Dai, PhD, American Board of Family Medicine, Lexington, Kentucky: Conceptualization, Specification, Testing

Craig Solid, PhD, Solid Research, LLC: Conceptualization, Specification, Testing

Jill Shuemaker, RN, CPHIMS, American Board of Family Medicine, Lexington, Kentucky: Conceptualization, Specification, Testing

Denise Pavletic, RD, MPH, American Board of Family Medicine, Lexington, Kentucky: Specification, Testing

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure? Current plans are to review/update annually, however; this may be adjusted as we learn more.

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

Ad.6 Copyright statement: None

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments: None