NATIONAL QUALITY FORUM

# MEASURE WORKSHEET

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

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

**Purple** text represents the responses from measure developers.

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

## Brief Measure Information

**NQF #:** 3227

**Corresponding Measures:**

**De.2. Measure Title:** CollaboRATE Shared Decision Making Score

**Co.1.1. Measure Steward:** The Dartmouth Institute for Health Policy & Clinical Practice

**De.3. Brief Description of Measure:** CollaboRATE is a patient-reported measure of shared decision making which contains three brief questions that patients, their parents, or their representatives complete following a clinical encounter. The CollaboRATE measure provides a performance score representing the percentage of adults 18 and older who experience a high level of shared decision making.

The measure was developed to be generic and designed so that it could apply to all clinical encounters, irrespective of the condition or the patient group. The measure asks the patient to evaluate the 'effort made' to inform, to listen to issues that matter to the patient, and to include those issues in choosing 'next steps'. The items were co-developed with patients using cognitive interview methods.

CollaboRATE is designed for use in routine health care delivery. The brevity and the ease of completion were purposeful, so the measure could be used as a performance metric for shared decision making.

**1b.1. Developer Rationale:** Measuring the level of shared decision making in the clinical encounter from the patient´s perspective is an important part of assessing healthcare quality and provider performance. CollaboRATE scores can provide important data to help facilitate quality improvement efforts across organizations.

Feedback from patients completing the survey demonstrates that they value being asked specifically about shared decision-making and being able to provide feedback on that topic. This patient feedback was obtained as part of a cognitive interview and pilot survey study (Elwyn 2013, doi: 10.1016/j.pec.2013.05.009). Participants in the pilot study (n=30) reported favorable views on the focus of the collaboRATE items, exemplified by participant quotations such as: "As many times as I have been here, I have never had a question like that. I think it's a damn good question."

**S.4. Numerator Statement:** Shared decision making; top-box scores represent the proportion of patients perceiving a high level of shared decision-making.

**S.6. Denominator Statement:** The denominator consists of all patients who complete the three CollaboRATE items. The denominator may include patients of any demographic or clinical background, as the measure is generic and applicable to a variety of clinical situations.

**S.8. Denominator Exclusions:** All patients are eligible to complete collaboRATE. Only incomplete collaboRATE responses should be excluded from the denominator.

**De.1. Measure Type:** Outcome: PRO-PM

**S.17. Data Source:** Instrument-Based Data

**S.20. Level of Analysis:** Clinician : Group/Practice

**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?** N/A

## Preliminary Analysis: New Measure

### Criteria 1: Importance to Measure and Report

#### 1a. Evidence

**1a. Evidence.** The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

**Evidence Summary**

- Brief background: This is a patient reported outcome performance measure (PRO-PM) of perception of shared decision making following a clinic encounter by patients aged 18+.
- Developer provided a logic model depicting the relationship between the content of a clinical encounter, the patient's experience of shared decision-making and their reporting of that experience
- Developer provided evidence of value and meaningfulness to the patient based on feedback from patients gathered through cognitive interviews and a pilot survey study (n=30).
- Empirical data demonstrating the relationship between a process to improve the outcome, developer cites a study of targeted shared decision-making interventions to improve performance on collaboRATE.

***Question for the Committee:***

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *Does the target population value the measured outcome and finds it meaningful?*

**Guidance from the Evidence Algorithm**

Measure assesses outcome (box 1) YES -> relationship between outcome and at least one healthcare action (box 2) YES -> PASS

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

## 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.

- Developer provides an analysis of data from 3 medical groups between April 2014 and October 2015 in NH, CA and MA totaling approximately 6,000 collaboRATE responses
    - Mean score and SD of 0.72 (0.09)
    - Score range of 0.68 to 0.86
- Developer also provides an analysis of data from the California Medical Group including approximately 31,000 respondents from primary and specialty site visits
    - Mean and SD: 0.60 (0.07)
    - Score range of 0.36 to 0.75
    - Scores by decile: 0.50, 0.54, 0.56, 0.58, 0.60, 0.61, 0.62, 0.64, 0.67, 0.71
- Results indicate a narrow range of performance in the first study, but a broader one in the second. Results indicate a moderate spread of performance indicated by the standard deviations.

**Disparities**

- Disparities were assessed between performance on English and Spanish forms of collaboRATE, showing similar scores.
- Developer notes that foreign language speakers that used an interpreter reported lower scores, suggesting that patients with less English familiarity that require an interpreter experienced poorer shared decision-making.

***Questions for the Committee:***

- Is the variation in performance sufficiently large to justify a performance measure?

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

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

**1a. Evidence**: *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.*

- Shared-Decision Making is a top priority for patient and family communities. Providers may additionally find the measure results useful.
- Pass - If nothing else, measuring Shared Decision Making (SDM) means it is more likely to take place!

**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?*

- My initial response was that the mean, range, and SD were high. I was expecting a larger performance gap, but still large gap considering it should happen in every clinical encounter. Demonstrates disparities in languages spoken, and I believe other subgroups of patient populations are likely experiencing large differences in the invitation for co-design. Would love to know how this performance gap compares to existing patient experience performance gaps.

- There were variations based on size of practices and English proficiency. Smaller practices may not take the time as often for SDM. Those with less English may not be getting the most out of SDM. Moderate GAP

## Criteria 2: Scientific Acceptability of Measure Properties

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

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

### Reliability

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

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

### Validity

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

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

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

**Evaluators:** NQF Scientific Methods Panel

Methods Panel Review (Combined)

**Methods Panel Evaluation Summary**:

This measure was reviewed by the Scientific Methods Panel and discussed on the call. A summary of the measure and the Panel discussion is provided below.

- **Ratings for reliability:** 5 moderate → Measure passes with MODERATE rating
    - Reliability testing included data element testing via internal consistency, as well as score-level testing via ICC and signal-to-noise (SNR) analysis (NOTE: that both data element and score-level reliability testing are required for instrument-based measures)
    - Data element testing
        - Internal consistency (using VA data)
            - Cronbach's alpha= 0.96 for inpatient respondents (n=767)
            - Cronbach's alpha= 0.97 for outpatient respondents (n=1,019)
    - Score-level testing
        - Data from California Medical Group Survey (153 groups, 31,265 responses from both primary and specialty outpatient care)
        - ICC=0.012; SNR: Median=0.720; $10^{th}$ percentile=0.634; $25^{th}$ percentile=0.678; $75^{th}$ percentile=0.746; $90^{th}$ percentile=0.757
            - Sample sizes in the clinician groups ranged from 31 (SNR=0.28) to 1133 (SNR=0.93), with an average of 204 responses per group
            - Used the reliability formula from Snijders & Bosker (1999)
- **Ratings for validity:** 1 high, 3 moderate, and 1 low → Measure passes with MODERATE rating

- Data element testing (i.e., validation of the instrument) was accomplished via face and content validity assessment, discriminative validity analysis to test discernment between levels of SDM, a sensitivity analysis of scores for a simulated encounter, and correlations of patient responses to four other tools (SDM-Q-9, 5-item Doctor Facilitation subscale of the Patient's Perceived Involvement in Care Scale, the Communication Assessment Tool (CAT), and the SHEP patient survey).
    - Face and content validity
        - Items were understood as intended by interview participants
        - Items, as understood by interview participants, covered all aspects of shared decision-making as modeled by experts in the field
    - Discriminative validity top score comparison (null hypothesis of no difference)
        - Comparison 1 (no SDM to low SDM): X2=24.9; p<0.001
        - Comparison 2 (low SDM to moderate SDM): X2=20.5; p<0.001
        - Comparison 3 (moderate SDM to high SDM): X2=4.7; p=0.03
    - Sensitivity analysis
        - CollaboRATE performance scores reflected SDM in 39% of clinical vignettes where all three dimensions of SDM were present
    - Correlations with other instruments
        - Concurrent validity with SDM Q-9: r=0.49; p<0.001
        - Concurrent validity with PICS-DFS: r=0.36; p<0.001
        - Concurrent validity with CAT among inpatients: r=0.84; p<0.001
        - Concurrent validity with CAT among outpatients: r=0.85; p<0.001
        - Concurrent validity with SHEP overall satisfaction among inpatients: r=0.74; p<0.001
        - Concurrent validity with SHEP overall satisfaction among outpatients: r=0.81; p<0.001
- Score-level testing accomplished by correlating clinician group scores with scores from two CG-CAHPS performance measures ("explanations are easy to understand" and "to what extent the doctor listens carefully").
    - Concurrent validity with CAHPS "explanations easy to understand" measure
        - In 92.8% of measured medical groups, correlations ≥ 0.60
        - In 69.3% of measured medical groups, correlations ≥ 0.70
        - Weakest correlation (one of 153 medical groups): r=0.4701; p<0.001
        - Strongest correlation (one of 153 medical groups): r=0.8971; p<0.001
    - Concurrent validity with CAHPS "listens carefully" measure
        - In 98.0% of measured medical groups, correlations ≥ 0.60
        - In 84.3% of measured medical groups, correlations ≥ 0.70
        - Weakest correlation (one of 153 medical groups): r=0.5805; p<0.001
        - Strongest correlation (one of 153 medical groups): r=0.8994; p<0.001
- In testing attachment, item on exclusions not completed (i.e., no analysis showing how many surveys not included in measure due to missing responses)
    - This information provided as part of missing data analysis (2b6)
    - In Single Group Primary Care Survey study, >99% of respondents completed all 3 CollaboRATE items; respondents were slightly older than non-respondents
- Risk adjustment – 2 risk factors (mode of survey administration; patient age) but **UNCLEAR** if interaction term is included or not
    - C-statistic=.6353 (95% CI 0.6178-0.6529)
    - Pseudo $R^2$=0.0354
    - Hosmer-Lemeshow statistic=773.46; p=0.1034

- Analysis of meaningful differences to indicate statistically significant differences between the three groups tested
  - Group 1 (OR 1.00)
  - Group 2 (OR 0.922, 95% CI 0.570-1.492
  - Group 3 (OR 1.759, 95% CI 1.216-2.545)
- Concerns
  - Unclear if all patients in a practice are given the survey (i.e., if patient significant involvement in shared decision-making is not expected or)
  - The 30-40% desired response rate is a little low, as is the 25 survey minimum
  - Concerns regarding lack of inclusion of social risk factors (unclear if dataset that was used to generate risk model was adequate)
  - Low sensitivity analysis result (39%)

***Questions for the Committee regarding reliability:***

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

***Questions for the Committee regarding validity:***

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

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

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

## Combined Methods Panel Scientific Acceptability Evaluation

**Measure Number: 3227**

**Measure Title: CollaboRATE Shared Decision Making Score**

### 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: Patient survey**

### Level of Analysis:

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

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

☐ **Integrated Delivery System** ☐☒ **Other MP#6:Individual interview participant / survey respondent / patient**

**Measure is:**

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

**MP#2**: Shared decision making measured by three questions. Numerator= top scores represent the proportion of patients perceiving a high level of shared decision-making. Denominator consists of all patients who complete the three CollaboRATE items. All patients are eligible to complete collaboRATE. Incomplete collaboRATE responses should be excluded from the denominator.

## 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.**

   **MP#4:**No concerns.

   **MP#6:**None.

## RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**

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

   **MP#2:**

   a. Cronbach's alha for internal consisten i cy. Values of 0.96 and 0.97 are quite good. with only three items it might suggest that the questions are redundant n the eyes/experience of the respondents.

   b. <u>*California Medical Group Survey:*</u> ICC of 0.012 across the full sample, 153 medical groups had a median reliability of 0.72. Reliabilities at the medical group level ranged from 0.28 in a medical group with only 31 patient collaboRATE responses to 0.93 in a medical group with 1133 patient collaboRATE responses.

   **Submission document:** Testing attachment, section 2a2.2

   **MP#3:**Reliability tests were conducted at both data element and measure score level. Group level reliabilities were derived from a hierarchical logistic regression model and were appropriately calculated. For data element reliability, internal consistency was calculated, however, it is not clear if it was based on top-box scaling or original scaling. It would be more appropriate to use the top-box scaling given how the measure score is calculated.

   **MP#5:** The methods used were traditional and appropriate, including Cronbach's alpha to establish reliability at the data element (individual patient survey) level and ICC and related reliability calculations to establish reliability at the measure score (clinic) level.

**MP#1:**The developer used Cronbach's alpha to evaluate internal consistency of the three items and a type of signal to noise analysis to evaluate practice level reliability. These approaches appear satisfactory.

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

   **Submission document:** Testing attachment, section 2a2.3

   **MP#6:**Appears that reliability is significantly impacted be the number of responses, if aggregating the results is acceptable then they demonstrated reliability using the deciles/quartile results.

   **MP#1:**Results support the reliability of the tool at the practice-level and the internal consistency of the items. The results indicate the the three measures within the measure 'hang together'. The practice level results indicate that the measure demonstrates reason

   **MP#4:** Test sample is adequate given the type of testing completed.

   **MP#3:**Internal consistency was very high with Cronbach's alpha of 0.96. Measure score reliabilities were moderate, mostly close to 0.6 or 0.7.
   **MP#5:**Results were generally acceptable, with high Cronbach's alpha results from a VA study and reliabilities generally exceeding .7 at the measure score level. The latter finding depends crucially on sample size, though, so any endorsement of the measure has to specify a minimum sample size.

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 <u>all</u> testing results):

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

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

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

    ☐ **Insufficient** (NOTE: Should rate <u>INSUFFICIENT</u> 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.**

    **MP#5:** Data element reliability seems to be high, and measure score reliability is generally acceptable, once a minimum sample size has been achieved. The testing form did not specify the minimum sample needed to achieve a reliability of at least .7, although a spreadsheet attachment could be used to do that. They developers could easily have done that in the measure testing form.

    **MP#1:**Internal consistency of the the three items is not a strong test of the reliability of the data elements.

    **MP#4:**Based on the amount of testing using various methods.

**MP#6:**Some concerns with the low reliability results of clinics with low response rates. It doesn't seem appropriate to imply that a clinic with 30 responses is comparable to a clinic with 300 responses

### VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

**MP#2**:

a. All patients eligible. Only those who provide incomplete or missing forms are excluded. Some concern regarding possible wide variation in response rate, and bias introduced by same. Note their target is 30-40% response rate.

**Submission document:** Testing attachment, section 2b2.

**MP#3:**No concern

**MP#6:**No exclusions

**MP#4:**Noted "no exclusions". Incomplete surveys are excluded from the denominator. No concerns.

**MP#5:**The developers do not identify any exclusions, but it would seem a little odd to include all patients in a survey when some of the patients do not need or desire any level of shared decision-making for the specific care they have received. Shared decision-making is good in general, but the concept does not apply to all interacctions (e.g., a visit to receive a routine immunization, or a visit to attend to a simple wrist sprain). Altuough users of the measure could assume that these "irrelevant" visits might be distributed evenly across clinics being evaluated, this is not necessarily so, particularly if the measure is being used to evaluate both primary care and specialty care groups.

**MP#1:**N/A

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

**MP#2**:

Requiring 3 of 3 top box ratings for a "1" (vs "0") score helps reduce ceiling effect but may fail to discriminate good from poor SDM. Only one "8" of "9" is a fail. Some concern there may be a false ceiling scored as a fail.

**Submission document:** Testing attachment, section 2b4.

**MP#3:**No concern

**MP#6:**See comments above

**MP#4:**No concerns.

**MP#5:**The measures seems able to distinguish between three clearly-defined levels of shared decision making in the simulation study, but it's less clear that the measure can distinguish among meaningful differences in more routine use. The developers don't seem to have a way to define what a "meaningful" difference would be, so it would be up to users to make that determination.

**MP#1:**The developers focused primarily on variation at the clinician level, and there is virtually no evidence of this measure's ability to identify meaning differences in practice/group performance (performance evaluated at 3 practices).

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

**MP#2**:

30-40% response rate concerns me on a measure like this, especially if there is variability around the response rate by clinic. Submitters suggest sequential rather than random or convenience sampling, and minimum of 25 responses. I might push that to over 50 and require 70-80% response rate.

**MP#6:**See comments above regarding different methods of administering
**MP#4:**NA
**MP#5:**None
**MP#1:**N/A

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

**Submission document:** Testing attachment, section 2b6.

**MP#3:**No concern

**MP#4:**High completion rate. No concerns.

**MP#1:**None.

**MP#5:None**

15. **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 **MP#1:** (N/A)

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

**MP#2**: Note social risk factors included in model but not retained in measure (see below)

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 NA

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**

**MP#4:**Not a lot of information about social risk factors other than they were examined. Risk factors include age and mode of administration.

**MP#6:**Agree with results showing only mode of suvey and patient age were the only factors impacting the survey results suppored by The Hosmer-Lemeshow analysis

**MP#2**: empirically tested the impact of the following social risk and other factors on CollaboRATE scores using mixed effects logistic regression analysis: survey administration mode; patient age; patient gender; number of health conditions; patient's primary language; patient race; and percent of residents below the federal poverty line in the patient's home zip code. Clinicians were included as a random effect in this model to account for clustering of patients by clinician. Results of the mixed effects logistic regression analysis showed that only patient age and survey administration mode had statistically significant associations with CollaboRATE scores, so they alone were retained in the risk adjustment model.

**MP#5:**The risk adjustment model only include two factors, and no social factors**.** The developers have apparently tested for the influence of other social factors (race, neighborhood SES) and found no effects.

**MP#3:** Risk-adjsutment approach was acceptable, however, it was based on a single-group primary care survey, it would be more desirable that additional tests be conducted using surveys from multiple group practices.

**MP#1:**Conceptually, this measure is one of the strongest candidates for inclusion of social risk factors given the compelling evidence on provision of care across race, ethnicity, education and gender as alluded to by the developer. My concern is that the data may not have been sufficiently available to test (data on frequency of these variables was not supplied) therefore they were not significant (model coefficients not provided). The c-statistic for the risk adjustment model is modest at .64.

**For cost/resource use measures ONLY:**

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

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

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

### VALIDITY: TESTING

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

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

    ☐☒ **Face validity**

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

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

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

    **MP#6:**Existing literature for face and content validity. Discriminate and concurrent validity results from Barr 2004 simulation study. VA study compared to CG-CAHPS communication items. California Medical Group Survey study, CollaboRATE performance scores were compared at the provider (medical group) level to scores on related CAHPS items using correlation analysis.

    **MP#2**:

    a. Face and content validity, discriminant (known groups) validity

    **Submission document: Testing attachment, section 2b2.2**

    **MP#4:**Good mix of methods used to establish validity.

    **MP#3:**Extensive validity tests were conducted, covering face and content validity, discriminative validity, concurrent validity and others.

    **MP#5:**Methods were generally acceptable for both levels of analysis – the use of the simulation study provides additional evidence of validity beyond that typically available for surveys of this type.

    **MP#1:**The developer briefly describes cognitive testing of wording in the data elements but there is limited information about the validity of the individual data elements from a conceptual (or empiric) perspective. The developer evaluated whether the summary measure score increased as the number of dimensions of SDM increased. The most compelling analysis of measure scores was the correlational analysis of the tool with other validated tools such as the CAHPS instrument, thereby evaluating criterion validity.

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

**MP#2**:

Chi-squared tests evaluated discriminant validity.

top score comparison 1 (no SDM to low SDM): $X^2=24.9$; $p<0.001$

top score comparison 2 (low SDM to moderate SDM): $X^2$=20.5; p<0.001

top score comparison 3 (moderate SDM to high SDM): $X^2$=4.7; p=0.03

Concurrent validity with SDM Q-9: r=0.49; p<0.001

Concurrent validity with PICS-DFS: r=0.36; p<0.001

VA study, CollaboRATE performance scores were compared to the Communication Assessment Tool (CAT) and overall satisfaction as measured in the SHEP survey using correlation analysis (Makoul 2007).

Concurrent validity with CAT among inpatients: r=0.84; p<0.001

Concurrent validity with CAT among outpatients: r=0.85; p<0.001

Concurrent validity with SHEP overall satisfaction among inpatients: r=0.74; p<0.001

Concurrent validity with SHEP overall satisfaction among outpatients: r=0.81; p<0.001

**Submission document: Testing attachment, section 2b2.3**

**MP#6:**All validity tests demonstrated statistically significant results.

**MP#4:**Appropriate

**MP#3:**Results from a couple of published studies demonstrated good face and content validity of CollaboRate. Simulation study also showed discriminative validity of CollaboRate performance score. CollaboRate performance scores were correlated with several relevant variables including previously validated related measures, satisfaction measures, and patient experience measures, correlations ranged from moderate to high.

**MP#5:**Results were generally acceptable, and show appropriate patterns of correlation both within the survey and with other related surveys (e,g, CAHPS in the California medical group study).

**MP#1:**Correlation of instrument scores with validated measures showed some evidence of measure score validity, although it isn't clear if risk adjusted scores were used for the CollaboRATE tool. Correlation with SDM tools was modest while correlations with concepts that were a bit orthogonal to SDM (i.e., satistfaction, communication assessment and CAHPS items assessing 'explainations were easy to understand') were higher suggesting relatively modest to low validity. As a PRO-PM, there was not evidence presented supporting empirical validitation of the CollaboRATE items – the face and content validity analysis focused on understand the items, not face validity.

22. **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)

23. **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)

24. **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.)

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

**MP#4:**May not have had a heterogenous sample to determine whether social risk factors apply.

**MP#6:**All methods of analysis supported validity

**MP#3:** Although many positive results on various validity testing were reported, it is concerning that "CollaboRATE performance scores reflected shared decision-making in 39% of clinical vignettes where all three dimensions of SDM were present."

**MP#2**: Although based on reliability level and concerns regarding missing data one might think I would rate moderate, I believe the intrinsic value, simplicity, and performance warrant a high rating

**MP#5:**The results were generally good – I held back the "high" rating to signal my concern about the issue raised above about exclusions. There are clearly differences among types of practices (oncology vs. primary care vs. urgent care) in terms of the appropriateness or necessity for shared decision-making. It is not clear how the variation observed from clinic to clinic reflects real differences in decision-making processes for the same underlying level of appropriateness or need vs. differences in case mix and patient interest in, or need for, shared decision-making.

**MP#1:**While there was some, albeit weak, evidence of score-level validity presented in the correlations with other instruments, there was no validation of the CollaboRATE items. The citations provided describe the model and its derivation but do not provide strong evidence of the validity of the items.

### ADDITIONAL RECOMMENDATIONS

26. **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.**

**MP#5:**The developers identify the measure as an Outcome measure and it is not. See my comments on various CAHPS measures for the full discussion on why this is a process measure and not an outcome measure.

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

**2a1. 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?*

- Seems that a simpler survey would be easier to consistently implement.
- Measure showed consistency.

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

- No
- Testing showed consistency

**2b2. Validity testing**: *Do you have any concerns with the testing results?*

- No

- Moderate.

**Validity- Threats to Validity**: *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?*

- N/A
- Moderate

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

- N/A
- SPM rated "moderate," noting it was unclear all persons in a practice received the survey (although SDM wouldn't necessarily occur in all encounters).

## Criterion 3. Feasibility

**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.

- Measure is available in both electronic and paper formats.
    - Electronic formats include web portal links via email, text message, or collected via a tablet.
- Developer notes the need to ensure patient confidentiality to protect the integrity of the scores.
- Measure is licensed under Creative Commons and is freely available.
- Developer notes in 4a2.2.1 that the questionnaire is completed by most patients in under 30 seconds.

***Questions for the Committee:***

- Is the burden to patients and providers to administer the survey outweighed by the value of the information gained?

**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?*

- No concerns regarding feasibility
- The SPM rated this Moderate. I would say "High." I think the survey itself could motivate and drive improved SDM.

## Criterion 4: Usability and Use

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

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

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

**Current uses of the measure**

**Publicly reported?** ☐ **Yes** ☒ **No**

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

**OR**

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

**Accountability program details**

- Blue Shield of California – Payment Program
- Pacific Business Group on Health – External benchmarking
- Right for me – External benchmarking
- Evaluating CollaboRATE – External benchmarking
- US Department of Veterans Affairs – Internal QI
- NQF has also received feedback from federal partners who have expressed interest in potentially using this measure for accountability purposes. Those plans have not been formalized to the knowledge of NQF staff.

**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**

**Additional Feedback:**

- Measure developer describes feedback in a single study conducted during development but does not describe how feedback is provided to clinicians in any of the accountability programs listed above.
- Paper-based in clinic surveys were viewed as a challenge due to staff administration burden.
- Automated administration did not have the same criticism.

***Questions for the Committee**:*

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

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

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

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

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

**Improvement results**

- Developers comments suggest that the measure has not been implemented long enough to have year over year data available for analysis

**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**

- None identified

**Potential harms**

- None identified

**Additional Feedback:** n/a

***Questions for the Committee:***

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

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

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

**4a. 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?*

- Would also like to know how clinicians responded to study implementation. My guess would be that it would be helpful information.
- The benefits strongly outweigh the burden of administration, IMO.

**4b. Usability**: *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.*

- My experience is that most clinicians find it difficult to reflect on how their patients experience SDM. Sometimes the experiences are very different, and this could be valuable feedback for establishing SDM as a common practice.
- No harms nor unintended consequences were identified.

## Criterion 5: Related and Competing Measures

**Related or competing measures**

The following measures are endorsed by NQF and could be considered related or competing:

- NQF 2962 Shared Decision Making Process

**Harmonization**

- Developer notes that the measure specifications are not completely harmonized as collaboRATE's target population is more inclusive than measure 2962 of Shared Decision Making Process, which focuses on patients undergoing specific surgical procedures. Instead, CollaboRATE allows for assessment of shared decision-making performance relevant to any type of health care encounter, context, or setting.

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

**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?*

- Seems that this one would be best in class compared to #2962 based on its' applicability across interactions.
- NA

## Public and Member Comments

**Comments and Member Support/Non-Support Submitted as of: June/13/2019**

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

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

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

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

nqf_evidence_CollaboRATE_7.1_for_Jan_2019-636915512450013820.docx

**1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last**

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

### 1a. Evidence (subcriterion 1a)

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

**Measure Title**: CollaboRATE Shared Decision Making Score

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

**Date of Submission**: 4/9/2019

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

Outcome

☐ Outcome:

☒ Patient-reported outcome (PRO):

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

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

☐ Process:

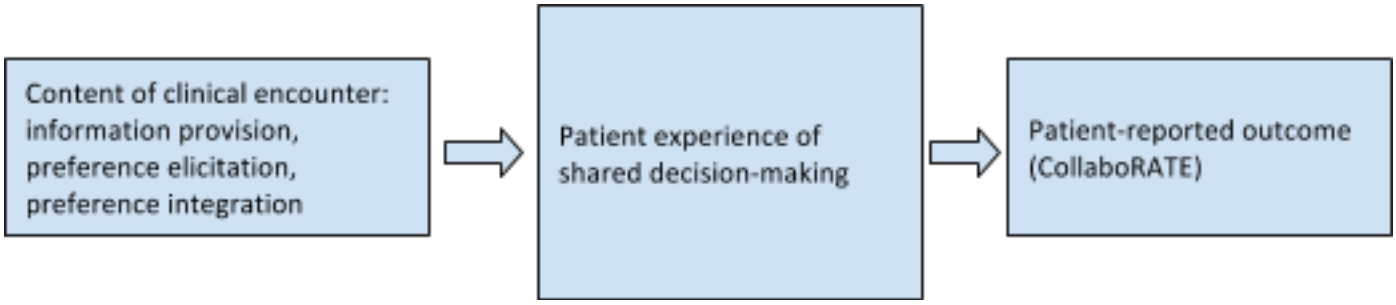
☐ Appropriate use measure:

☐ 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.

This PRO directly measures the three key elements of the process of shared decision-making, namely: Information provision; Preference elicitation; and Preference integration (Elwyn G, Barr PJ, Grande SW, Thompson R, Walsh T, Ozanne EM. 2013. Developing CollaboRATE: a fast and frugal patient-reported measure of shared decision making in clinical encounters. doi:10.1016/j.pec.2013.05.009).




**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.)

Feedback from patients completing the survey demonstrates that they value being asked specifically about shared decision-making and being able to provide feedback on that topic. This patient feedback was obtained as part of a cognitive interview and pilot survey study (Elwyn 2013, doi: 10.1016/j.pec.2013.05.009). Participants in the pilot study (n=30) reported favorable views on the focus of the collaboRATE items, exemplified by participant quotations such as: "As many times as I have been here, I have never had a question like that. I think it's a damn good question."

****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.**

CollaboRATE has been shown to respond to clinical interventions to improve shared decision-making. In Tai-Seale's 2016 cluster-randomized trial of a shared decision-making intervention called OpenComm (doi: 10.1377/hlthaff.2015.1398), clinics randomized to the intervention arm had significantly higher collaboRATE scores than clinics which were randomized to the usual care arm (OR 1.523; 95% CI 1.026-2.259). This evidence demonstrates that collaboRATE can be influenced by targeted shared decision-making interventions, and that it is possible for providers to improve their collaboRATE scores through use of such interventions.

**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:**<br>• **Title**<br>• **Author**<br>• **Date**<br>• **Citation, including page number**<br>• **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:<br>• Quantity – how many studies?<br>• 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.

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

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

## 1b. Performance Gap

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

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

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

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

Measuring the level of shared decision making in the clinical encounter from the patient´s perspective is an important part of assessing healthcare quality and provider performance. CollaboRATE scores can provide important data to help facilitate quality improvement efforts across organizations.

Feedback from patients completing the survey demonstrates that they value being asked specifically about shared decision-making and being able to provide feedback on that topic. This patient feedback was obtained as part of a cognitive interview and pilot survey study (Elwyn 2013, doi: 10.1016/j.pec.2013.05.009). Participants in the pilot study (n=30) reported favorable views on the focus of the collaboRATE items, exemplified by participant quotations such as: “As many times as I have been here, I have never had a question like that. I think it’s a damn good question.”

**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.*

Three Group Primary Care Survey

In a study of 3 medical groups, collaboRATE performance scores were compared across groups. All consecutive adult patients visiting the three primary care groups during their respective survey periods (between April 2014-October 2015) were eligible to complete CollaboRATE. One group was in Lebanon, NH with a patient population of 16,000; a second group was in Palo Alto, CA with a patient population of 13,000; the third group was in Chelsea, MA with a patient population of 14,000. Across the three sites, 5974 patient collaboRATE responses were collected: 4421 in group 1, 323 in group 2, and 1230 in group 3. Performance score data, shown in the table below, demonstrate a gap in collaboRATE performance scores between medical groups.

Mean score: 72%

Standard deviation: 9

Score range (min-max): 68%-86%

Interquartile range: --

Scores by decile: --

Number of measured entities: 3

Number of patient reports: 5974

California Medical Group Survey

California Medical Group data were collected as part of the Patient Assessment Survey (PAS) administered by the Pacific Business Group on Health. This survey is comprised of the Clinician-Group CAHPS instrument plus CollaboRATE and other items of interest. CollaboRATE performance scores were compared across medical groups. The California Medical Group Survey included 31,265 respondents, with 16,627 answering based on a primary care visit and 14,638 answering based on a specialty care visit. Participants are from a random sample of adult and pediatric visits by insured patients across 153 California medical groups, including both primary and specialty outpatient care. Performance score data, shown in the table below, demonstrate a gap in collaboRATE performance scores between medical groups.

Mean score: 60%

Standard deviation: 7

Score range (min-max): 36%-75%

Interquartile range: 8

Scores by decile: 50%, 54%, 56%, 58%, 60%, 61%, 62%, 64%, 67%, 71%

Number of measured entities: 153

Number of patient reports: 31,265

**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.**

CollaboRATE was created with substantial input from patients and laypeople to address a lack of existing measurement tools in the area of patient engagement and shared decision-making that: 1) can be administered close to the time of the encounter to enhance recall; 2) are brief and easy to complete to reduce burden of both data collection and patient response; 3) specific to the core constructs of shared decision-making, namely information sharing, preference elicitation, and preference integration (Elwyn G, Barr PJ, Grande SW, Thompson R, Walsh T, Ozanne EM. 2013. Developing CollaboRATE: a fast and frugal patient-reported measure of shared decision making in clinical encounters. doi:10.1016/j.pec.2013.05.009; Barr PJ, Thompson R, Walsh T, Grande SW, Ozanne EM, Elwyn G. 2015. The Psychometric Properties of CollaboRATE: A Fast and Frugal Patient-Reported Measure of the Shared Decision-Making Process. doi:10.2196/jmir.3085).

Testing has shown variation in CollaboRATE performance scores across clinicians, with scores ranging from 61% to 81% across three clinics and from 42% to 99% across clinicians.

**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.*

Preliminary research in Chelsea, MA, an area where the population has low socio-economic status and high levels of immigration and diversity, has shown similar scores between those patients completing CollaboRATE in English (CollaboRATE score=86%, n=624) and those completing it in Spanish (CollaboRATE score=86%, n=606). However, among respondents to the English version (n=586), scores were higher among patients who did not have a language interpreter present during their medical appointments (CollaboRATE score=88%, n=517) than among those who had a language interpreter present (CollaboRATE score=74%, n=69), suggesting that patients with less English familiarity who required an interpreter experienced poorer shared decision-making.

**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**

## 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.)*

http://www.glynelwyn.com/collaborate-measure.html

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

This is not an eMeasure **Attachment:**

**S.2b. Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

No data dictionary **Attachment:**

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

«instrument_based» **Attachment:** «instrument_based_provided»

**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.

«responder»

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

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

**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).*

Shared decision making; top-box scores represent the proportion of patients perceiving a high level of shared decision-making.

**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 consists of those cases (i.e. patient responses) where perfect scores are given on all three CollaboRATE items; cases with perfect scores are coded '1', whereas all other patient scores are coded '0' in a dichotomous top score outcome variable.

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

The denominator consists of all patients who complete the three CollaboRATE items. The denominator may include patients of any demographic or clinical background, as the measure is generic and applicable to a variety of clinical situations.

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

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

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

CollaboRATE is applicable to all patients; the denominator therefore consists of all complete responses.

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

All patients are eligible to complete collaboRATE. Only incomplete collaboRATE responses should be excluded from the denominator.

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

Exclude from the denominator any cases in which there are missing responses on any of the three collaboRATE items.

**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.)*

We do not stratify by patient or provider level characteristics, although there may be analytic interest in these variables. If responses are collected for patients of all ages, it may be appropriate to stratify by pediatric and adult patients.

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

Statistical risk model

If other:

**S.12. Type of score:**

Rate/proportion

If other:

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

Better quality = Higher score

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

To calculate CollaboRATE Performance Score:

Exclude cases (i.e. patient survey responses) where a response to one or more of the CollaboRATE questions is missing. Code each case as either ´1´, if the response to all three CollaboRATE items was 9, or ´0´ if the response to any of the three CollaboRATE items was less than 9. To case-mix adjust scores, conduct logistic regression analysis with the binary collaboRATE score outcome as the dependent variable and independent variables including patient age and patient gender; predict probabilities at the medical group level based on this model. These probabilities are the CollaboRATE performance scores for each medical group. Higher scores represent more shared decision making. This number also corresponds to the case-mix adjusted proportion of patients who perceive ´gold standard´ shared decision making.

**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.

Due to collaboRATE's broad applicability to a wide variety of clinical contexts, the sample can consist of all consecutive patients. Consecutive patient samples are more feasible for timely implementation in routine practice than random samples, while also minimizing selection bias as compared to convenience sampling. In determining an optimal sampling strategy for a given use of collaboRATE, and especially when considering consecutive sampling, seasonal and time-dependent trends that affect patient flow should be taken into account. A minimum of 25 responses per unit of analysis is required (Chisholm & Askham, 2006, What do you think of your doctor: a review of questionnaires for gathering patients' feedback on their doctor).

CollaboRATE was designed to be applicable to all clinical settings and any type of healthcare decision - from minor and sometimes implicit decisions like continuing with an already-established treatment plan to major decisions such as designing a brand new treatment plan. In developing the measure together with target end-users, we confirmed the measure was usable and understandable in a wide variety of clinical contexts and circumstances for assessing the core components of shared decision-making, i.e. 1) information provision; 2) preference elicitation; and 3) preference integration.

Proxy responses are allowed depending on the specific target population. A proxy version of the measure for parents can be found at: http://www.glynelwyn.com/uploads/2/4/0/4/24040341/collaborate_forparents_v4_1.pdf. A proxy version for other individuals acting on behalf of patients can be found at: http://www.glynelwyn.com/uploads/2/4/0/4/24040341/collaborate_forproxies_v2_1.pdf.

**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.

At survey administration, we aim for a 30-40% response rate. If administered within the clinic setting, surveys should be administered in a private area to ensure confidentiality of responses and minimize potential influence of clinic staff on survey responses. To facilitate survey administration, electronic (e.g. text message, telephone) and online (e.g. patient portal) methods are recommended.

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

*If other, please describe in S.18.*

Instrument-Based Data

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

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

Multiple modes of data collection, including: Paper-based, tablet, text messages (SMS), online patient portal, and interactive voice response (automated telephone calls). Tool is adaptable for use across multiple data collection modalities.

Measure is available in English, Spanish, Dutch, Norwegian, Danish, and French.

**S.19. Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

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

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

Clinician : Group/Practice

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

Inpatient/Hospital, Outpatient Services

If other:

**S.22. <u>COMPOSITE Performance Measure</u>** - 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_7.1_CollaboRATE_final_for_Jan_2019.docx,Appendix_1_-_Stata_output_for_signal-to-noise_analysis_final_for_Jan_2019.docx

**2.1 <u>For maintenance of endorsement</u>**

*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 <u>For maintenance of endorsement</u>**

*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 <u>For maintenance of endorsement</u>**

*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**: CollaboRATE Shared Decision Making Score
**Date of Submission**:

**Type of Measure:**

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

**1. DATA/SAMPLE USED FOR <u>ALL</u> 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. <u>If there are differences by aspect of testing</u>,(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: Instrument (Patient survey) | ☒ other: Interviews; Clinical simulation survey; Patient survey |

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

**1.3. What are the dates of the data used in testing**? Cognitive interviews: 2012; Online Simulation Survey: January-February 2013; Single Group Primary Care Survey: April 2014-October 2015; Three Group Primary Care Survey: April 2014-October 2015; Veterans Administration Survey: October 2013-September 2014; California Medical Group Survey: December 2016-March 2017

**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: Individual interview participant / survey respondent / patient |

**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*)

Cognitive Interviews

Participants in the cognitive interview study were individuals recruited from the public areas of an academic medical center to provide feedback on the face and content validity of the CollaboRATE items and response anchors. Twenty-seven individuals completed interviews, of whom 15 were female, 8 were ages 18-44, 11 were ages 45-64, and 8 were age 65 or older. Study described in detail in: Elwyn et al. 2013. doi:10.1016/j.pec.2013.05.009.

Clinical Simulation Survey

Individuals in the clinical simulation study were recruited from an online survey panel for participation in a simulation study using clinical vignettes (n=1341). Sampling quotas ensured that respondents represented the United States population with regard to age, gender, and race. Study described in detail in: Barr PJ et al. 2014. The psychometric properties of CollaboRATE: a fast and frugal patient-reported measure of the shared decision making process. doi:10.2196/jmir.3085.

Single Group Primary Care Survey

In this patient survey study, all consecutive patients visiting three primary care clinical teams within a single primary care practice were eligible to complete CollaboRATE and we obtained responses from 4421 of 17568 patients. Study described in detail in: Barr PJ et al. 2017. Evaluating CollaboRATE in a clinical setting: analysis of mode effects on scores, response rates, and costs of data collection. doi:10.1136/bmjopen-2016-014681.

Three Group Primary Care Survey

In a study at 3 clinical groups, CollaboRATE performance scores were compared across groups/practices. All consecutive adult patients visiting three primary care practices during their respective survey periods were eligible to complete CollaboRATE. One group was in Lebanon, NH with a patient population of 16,000; a second group was in Palo Alto, CA with a patient population of 13,000; the third group was in Chelsea, MA with a patient population of 14,000. Study described in detail in: Forcino RC et al. 2018. Using CollaboRATE, a brief patient-reported measure of shared decision making: results from three clinical settings in the United States. doi:10.1111/hex.12588.

Veterans Administration Survey

VA data were collected from four VA Office of Patient-Centered Care and Cultural Transformation Centers of Innovation, as well as from matched comparison facilities. Survey data includes responses from both inpatients and outpatients across medical specialties.

California Medical Group Survey

California Medical Group data were collected as part of the Patient Assessment Survey (PAS) administered by the Pacific Business Group on Health. This survey is comprised of the Clinician-Group CAHPS instrument plus CollaboRATE and other items of interest. The PAS is the nation's largest system for evaluating and publishing physician group ratings based on the patient's experience, with all California medical groups participating in commercial health insurance eligible for participation. Survey responses are from a random sample of adult and pediatric visits by insured patients across California's medical groups, including both primary and specialty outpatient care. CollaboRATE performance scores were compared across medical groups.

**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*)

Cognitive Interviews

Participants in the cognitive interview study included 27 individuals recruited from public areas of an academic medical center, of whom 15 were female, 8 were ages 18-44, 11 were ages 45-64, and 8 were age 65 or older. Purposeful sampling ensured a diverse group of participants with regard to age and gender.

Clinical Simulation Survey

Simulations using clinical vignettes included a sample of 1341 individuals recruited from an online survey panel. 54% of the 1341 participants were female, 81% were white, and 53% were age 45 or older (Barr 2014). The non-probability sample was recruited by Survey Sampling International, who employed quotas to ensure participation across age groups and genders.

Single Group Primary Care Survey

The patient survey included 4421 patients recruited from a primary care practice. Respondents were representative of the patient population with regard to gender (66% female) and had a mean age of 50 (Barr 2017). All consecutive adult patients visiting participating clinical teams over the 15-month study period were

eligible to participate. Over the 15-month study period, five modes of patient survey administration were used for three months each: paper questionnaire in-clinic; online EHR portal survey; text message (SMS) survey; automated phone call (IVR) survey; in-clinic tablet computer survey.

Three Group Primary Care Survey

This patient survey included 5974 patient responses across three primary care clinical groups: 4421 in group 1 (see "Single Group Primary Care Survey" above), 323 in group 2, and 1230 in group 3. All consecutive patients visiting participating clinical teams over the groups' respective study periods were eligible to participate.

Veterans Administration Survey

The VA patient survey included 767 inpatients and 1019 outpatients who received care in VA facilities. CollaboRATE was administered to this sample as a component of a routine patient survey. Available participant demographic data is limited for this sample.

California Medical Group Survey

The California Medical Group Survey included 31,265 respondents, with 16,627 answering based on a primary care visit and 14,638 answering based on a specialty care visit. Participants are from a random sample of adult and pediatric visits by insured patients across 153 California medical groups, including both primary and specialty outpatient care. 62% of participants were female, 60% were white, and 39% were between the ages of 55 and 64. 82% of participants self-reported good, very good, or excellent health.

**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**.

Literature Review and Cognitive Interviews

Existing literature describes the theoretical background on which CollaboRATE is based, including SDM models developed by experts including Elwyn 2012 (doi: 10.1007/s11606-012-2077-6) and Makoul 2006 (doi:10.1016/j.pec.2005.06.010). Face and content validity of this patient-reported measure was tested through cognitive interviews with members of the measure's target population - specifically, members of the public within an academic medical center who may have future experience as a patient or caregiver. This sample is described in detail in sections 1.5 and 1.6 above.

Clinical Simulation Survey

The clinical simulation study featured clinical vignettes, with varying levels of shared decision making (the construct that CollaboRATE measures), presented in a web-based survey to individuals who were recruited from an online survey panel and were not required to have recently attended a clinical visit. The simulated nature of this study provided an opportunity to collect extensive participant data and perform psychometric testing. Our testing of concurrent validity, sensitivity, discriminative validity uses this data. This sample is described in detail in sections 1.5 and 1.6 above.

Single Group Primary Care Survey

This study allowed for examination of the effect of case mix on CollaboRATE scores, which informed the risk adjustment model. This sample is described in detail in sections 1.5 and 1.6 above.

Three Group Primary Care Survey

This study allowed for testing of differences in CollaboRATE performance at both the individual clinician and group levels. This sample is described in detail in sections 1.5 and 1.6 above.

Veterans Administration Survey

The VA study provided concurrent validity information related to the relationship between CollaboRATE performance and overall patient satisfaction (as measured with SHEP questionnaire) and the relationship between performance on CollaboRATE and Makoul's (2007) Communication Assessment Tool. The VA study also provided data on internal consistency and reliability of the CollaboRATE measure. Finally, this study

demonstrated versatility of the CollaboRATE measure by collecting data from both inpatient and outpatient clinical encounters. This sample is described in detail in sections 1.5 and 1.6 above.

California Medical Group Survey

The California Medical Group study demonstrated versatility of the CollaboRATE measure by collecting data in both primary and specialty care settings. It also provided reliability data at the medical group level. Finally, it provided concurrent validity data at the medical group level, allowing for correlation analysis between CollaboRATE scores and related CAHPS items. This sample is described in detail in sections 1.5 and 1.6 above.

**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.

Cognitive Interviews
In the cognitive interview study, sociodemographic variables included gender, age, educational attainment, and employment status.
Clinical Simulation Survey
In the clinical simulation study, sociodemographic variables included patient-reported gender, age, educational attainment, ethnicity, race, language spoken at home, illness and/or disability status, and recent health care experience.
Single Group Primary Care Survey
In the Single Group Primary Care Survey study, sociodemographic variables included patient-reported gender and age. Further, demographic data were available through the clinic's electronic medical record, including diagnostic codes, race, and primary language. The proportion of residents below the federal poverty line within the patient's home zip code was included in analysis as a patient community characteristic.
Three Group Primary Care Survey
In the Three Group Primary Care Survey, sociodemographic variables collected included patient-reported gender and age.
Veterans Administration Survey
No social risk factor data were analyzed in this study.
California Medical Group Survey
Social risk factors analyzed in this study, namely as part of the case mix adjustment strategy, include mode of survey administration and patient age.

---

**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 account for ceiling effects common to patient-reported experience measures, CollaboRATE performance scores are top box scores; CollaboRATE score is therefore a dichotomous variable coded '1' where the highest possible response is given for each of the three CollaboRATE items or '0' if any of the three items receives a non-optimal score. The proportion of top scores for a measured entity represents that entity's CollaboRATE performance score.

**Reliability of provider profiling/differentiating provider performance:** In the *California Medical Group Survey* study, mixed effects regression analysis estimated medical group-level random effect variance. As described in Adams' (2009) Reliability of Provider Profiling tutorial we used the *California Medical Group Survey* dataset to conduct mixed effects logistic regression analysis and subsequently calculate the intraclass correlation (ICC). We then used the reliability formula from Snijders & Bosker's (1999) text (reproduced below) to calculate reliabilities for each medical group, where ρ was the case-mix adjusted ICC and $n$ was the sample size for the given medical group. We calculated reliabilities for each of the 153 medical groups and presented the median value, following the example of prior literature (Scholle 2008).

Our steps included:

*-Conducted mixed effects logistic regression analysis using Stata version 13.1 to estimate medical group-level ICC. The dependent variable was the patient-level collaboRATE score and the medical group ID was included as a random effect; patient age and gender were included as fixed effects. See attached Appendix 1 for full model specification including Stata code and output.*

*-Used the reliability formula included in Snijders & Bosker's (1999) multilevel analysis textbook (reproduced below) to calculate a reliability estimate for each of the 153 included medical groups using the case mix-adjusted ICC and group-level sample sizes.*

$$reliability = \frac{n * \rho}{1 + (n - 1)\rho}$$

*Our 153 reliability calculations incorporated the following values:*

ρ *= case-mix adjusted ICC*

*n = sample size for the given medical group*

*-Of the 153 reliability estimates calculated in step 2, we present the median value as demonstrated in Scholle's 2008 'Benchmarking Physician Performance: Reliability of Individual and Composite Measures.' We also present the 10th, 25th, 75th, and 90th deciles/quartiles.*

**Internal consistency:** In a sample of Veterans Administration inpatients and outpatients, Cronbach's alpha was used to assess internal consistency and reliability of CollaboRATE at the patient-level data element.

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

**Reliability of provider profiling/Differentiating individual and group performance:** *California Medical Group Survey:* With an ICC of 0.012 across the full sample, medical groups (n=153) had a median reliability of 0.720 (n=204). Reliabilities at the medical group level ranged from 0.28 in a medical group with only 31 patient collaboRATE responses to 0.93 in a medical group with 1133 patient collaboRATE responses. See attached spreadsheet (Appendix 2) for reliability computations for each of the 153 medical groups.

| | **Reliability estimate** |
|---|---|
| First decile | 0.634 |
| First quartile | 0.678 |
| Median | 0.720 |
| Third quartile | 0.746 |
| Ninth decile | 0.757 |

**Internal consistency:** Data collected in Veterans Administration settings showed high reliability and internal consistency at the patient-level data element, with Cronbach's alpha of 0.96 among inpatient respondents and 0.97 among outpatients.

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

**Reliability of provider profiling/Differentiating provider performance:** According to Adams' (2009) Reliability of Provider Profiling, "psychometricians use a rule of thumb of 90 percent [0.9] for drawing conclusions about individuals. Lower levels (70-80 percent) are considered acceptable for drawing conclusions about groups" (Adams 2009). Therefore, given that the average medical group reliability derived from the *California Medical Group Survey* met the 70 percent (0.7) reliability threshold, our results indicate acceptable reliability in differentiating CollaboRATE performance at the medical group level.

**Internal consistency:** Data collected in the Veterans Administration Survey showed high reliability and internal consistency at the patient-level data element, with Cronbach's alpha of 0.96 among inpatient respondents and 0.97 among outpatients.

---

**2b1. VALIDITY TESTING**

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

**Face validity and content validity:** Existing literature describes the theoretical background on which CollaboRATE is based, including SDM models developed by experts including Makoul 2006 (doi:10.1016/j.pec.2005.06.010), Elwyn 2012 (doi:10.1007/s11606-012-2077-6), and Elwyn 2017 (doi:10.1136/bmj.j4891). Face validity testing described in Elwyn (2013) used one-on-one cognitive interviews and qualitative analysis to assess whether identified experts - members of the public who may seek health care in the future - understood the meaning of the items in a way that is consistent with their intended meaning. Interview questions included, "In your own words, what do you think the question is asking?", "Do the words in the question make sense?", and "What does the term 'how much effort' mean to you?". Cognitive interviewing is defined by Willis (2013) as an evidence-based, qualitative method specifically designed to investigate whether a survey question - whether attitudinal, behavioral, or factual in nature - fulfills its intended purpose (Willis 2013, doi: 10.4300/JGME-D-13-00154.1), in this case, meaning whether or not it measures the core components of shared decision-making established by experts in prior literature (Elwyn 2012, Makoul 2006).

To ensure that variation observed at the provider level is attributable to differences in SDM performance, careful attention throughout the measure development process was given to ensuring face validity of included items and their ease of interpretation by the target audience of measure respondents. This process ensured interpretability of each item as a direct assessment of provider SDM performance. As a patient-reported experience measure, and unlike a biological outcome measure, collaboRATE's wording and response scale were developed in partnership with patients to be tailored to specifically measure provider performance across patient groups. Elwyn's 2013 article details CollaboRATE development, focusing on use of language that is understandable and interpretable to the measure's target respondent audience, including a focus on provider effort and avoidance of terms found difficult to interpret such as 'decision' and 'problem'.

**Discriminative validity:** Discriminative validity testing used population-level scores captured in the simulation study (Barr 2014) and represents testing of the CollaboRATE performance score. Discriminative validity testing detailed in Barr (2014) assessed whether CollaboRATE top scores increased as the number of dimensions of shared decision-making portrayed in the simulated clinical encounters increased; chi-squared tests and between-groups/Welch's t-tests were used. In this study there were four vignettes; one depicted a high level of SDM, one depicted a moderate level of SDM, one depicted a low level of SDM, and one depicted no SDM. These summary scores were assigned to clinical vignettes based on the number of SDM dimensions they portrayed, including 1) information exchange about treatment options; 2) elicitation of patients' preferences pertaining to treatment options; and 3) integration of patients' preferences into next steps. A 'high' score of three means that all three dimensions were present in the clinical vignette. The three discriminative validity comparisons listed in 2b1.3 below represent: 1) no SDM vs. low SDM; 2) low SDM vs. moderate SDM; 3) moderate SDM vs. high SDM. This testing assumes a null hypothesis of no difference between comparison groups.

**Sensitivity:** In this context, sensitivity testing involves calculating the number of 'true positives', or cases in which CollaboRATE detects shared decision-making when it is objectively present in a clinical encounter. This calculation is based on data from the Online Simulation Survey study, in which there was by design an objective measure of shared decision-making to which we compared CollaboRATE scores.

**Concurrent validity:** In the Online Simulation Survey study (Barr 2014), concurrent validity testing evaluated the strength of the association between CollaboRATE performance scores and scores on previously-validated related measures (SDM-Q-9 and 5-item Doctor Facilitation subscale of the Patient's Perceived Involvement in Care Scale [PICS]) of shared decision-making and clinical communication; point-biserial correlations were used. The SDM-Q-9 (Kriston 2010) is a 9-item measure of shared decision-making with 6-option Likert-type response scale designed for use in research studies, while the Doctor Facilitation subscale of the Perceived Involvement in Care scale (Lerman 1990) measures the extent to which doctors facilitate patient participation in healthcare encounters, a closely related but separate construct to SDM. As such, these two measures are appropriate for comparison in CollaboRATE's concurrent validity assessment.

In the Veterans Administration Survey study, we examined correlations between CollaboRATE and the Communication Assessment Tool developed by Makoul (2007) to measure the related construct of clinician-patient communication. We also examined correlations between CollaboRATE and overall patient satisfaction as measured in the VA's routine SHEP patient survey.

In the California Medical Group Survey study, we examined correlations between average CollaboRATE scores and CG-CAHPS communication items, namely those CAHPS items asking to what extent the doctor's explanations are easy to understand and to what extent the doctor listens carefully. We analyzed these correlations by medical group (n=53 groups) to establish provider-level concurrent validity.

**External validity:** Data from 1) primary care settings in three distinct medical groups, 2) Veterans Administration inpatient and outpatient settings, and 3) primary and specialty care settings in medical groups across California demonstrate versatility of the CollaboRATE measure and its applicability in a range of contexts.

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

**Discriminative validity:** Chi-squared tests evaluated discriminative validity.

Discriminative validity top score comparison 1 (no SDM to low SDM): $X^2=24.9$; $p<0.001$

Discriminative validity top score comparison 2 (low SDM to moderate SDM): $X^2=20.5$; $p<0.001$

Discriminative validity top score comparison 3 (moderate SDM to high SDM): $X^2=4.7$; $p=0.03$

Interpretation of these results is detailed in section 2b1.4 below.

**Sensitivity:** In the Online Simulation Survey study, CollaboRATE performance scores reflected shared decision-making in 39% of clinical vignettes where all three dimensions of SDM were present. More research is needed in non-simulated settings to compare CollaboRATE scores to objective/observer measures of SDM, as

CollaboRATE scoring takes into account ceiling effects resulting from common patient response biases; lower measure sensitivity in a simulated setting is therefore anticipated.

**Concurrent validity:** In the Online Simulation Study, point-biserial correlations compared CollaboRATE to previously-validated measures (SDM-Q-9 and PICS) to evaluate concurrent validity.

Concurrent validity with SDM Q-9: r=0.49; p<0.001

Concurrent validity with PICS-DFS: r=0.36; p<0.001

In the VA study, CollaboRATE performance scores were compared to the Communication Assessment Tool (CAT) and overall satisfaction as measured in the SHEP survey using correlation analysis (Makoul 2007).

Concurrent validity with CAT among inpatients: r=0.84; p<0.001

Concurrent validity with CAT among outpatients: r=0.85; p<0.001

Concurrent validity with SHEP overall satisfaction among inpatients: r=0.74; p<0.001

Concurrent validity with SHEP overall satisfaction among outpatients: r=0.81; p<0.001

In the California Medical Group Survey study, CollaboRATE performance scores were compared at the provider (medical group) level to scores on related CAHPS items using correlation analysis.

*Concurrent validity with CAHPS "explanations easy to understand" by medical group*

Weakest correlation at medical group level (one of 153 medical groups): r=0.4701; p<0.001

Strongest correlation at medical group level (one of 153 medical groups): r=0.8971; p<0.001

In 92.8% of measured medical groups, correlations between CollaboRATE and CAHPS "explanations easy to understand" item were equal to or exceeded 0.60. In 69.3% of measured medical groups, correlations between CollaboRATE and CAHPS "explanations easy to understand" item were equal to or exceeded 0.70.

*Concurrent validity with CAHPS "listens carefully" by medical group*

Weakest correlation at medical group level (one of 153 medical groups): r=0.5805; p<0.001

Strongest correlation at medical group level (one of 153 medical groups): r=0.8994; p<0.001

In 98.0% of measured medical groups, correlations between CollaboRATE and CAHPS "listen carefully" item were equal to or exceeded 0.60. In 84.3% of measured medical groups, correlations between CollabORATE and CAHPS "listen carefully" item were equal to or exceeded 0.70.

Interpretation of these results is detailed in section 2b1.4 below.

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

**Face and content validity:** CollaboRATE shows strong face validity as the items were understood as intended by interview participants. CollaboRATE shows strong content validity as the items, as understood by interview participants, covered all aspects of shared decision-making as modeled by experts in the field (Elwyn 2012; Makoul 2006).

**Discriminative validity:** CollaboRATE shows strong discriminative validity in discerning between different levels of shared decision-making, as the null hypothesis of no difference between comparison groups was rejected at p<0.05 for each comparison made (i.e. no SDM vs. low SDM; low SDM vs. moderate SDM; moderate SDM vs. high SDM).

**Sensitivity:** In the Online Simulation Survey study, CollaboRATE performance scores reflected shared decision-making in 39% of clinical vignettes where all three dimensions of SDM were present. More research is needed in non-simulated settings to compare CollaboRATE scores to objective/observer measures of SDM, as CollaboRATE scoring takes into account ceiling effects resulting from common patient response biases; lower measure sensitivity in a simulated setting is therefore anticipated.

**Concurrent validity:** A strong correlation demonstrating strong concurrent validity should exceed 0.60. In the Online Simulation Survey, CollaboRATE top score shows moderate concurrent validity with established measures of shared decision-making including SDM-Q-9 and PICS-DFS. In the VA Survey study, we observe

strong correlation between CollaboRATE and CAT with correlation coefficients of 0.84 and 0.85 and strongly significant at p<0.001. We also observe strong correlation between CollaboRATE and SHEP overall satisfaction with correlation coefficients of 0.81 and 0.74 for outpatients and inpatients respectively, which were strongly significant at p<0.001. In the California Medical Group Survey study, we found high concurrent validity between CollaboRATE scores at the medical group level, with correlations between CAHPS communication items exceeding the threshold for strong correlation (0.60) in more than 90% of all 153 medical groups.

---

**2b2. EXCLUSIONS ANALYSIS**

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

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

**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* 2b4.**

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 2 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.**

Logistic regression model with robust standard errors and risk factors/independent variables including: mode of survey administration; patient age.

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

N/A

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

**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?
Conceptually, there is reason to suspect a linkage between shared decision-making as measured by CollaboRATE and social risk factors such as patient education, primary language, race, ethnicity, and income.

Shared decision-making itself can be impacted by lack of language concordance as well as perceived power imbalances between patient and provider, and prior research has explored survey response biases associated with diverse respondent characteristics in healthcare settings. The collaborative deliberation model of clinical communication, developed by Elwyn et al. (2014), is designed to address the disempowerment that is inherent to patient-professional interactions and which can be exacerbated by the risk factors discussed above.
For these reasons, in developing our statistical risk model, we empirically tested the impact of the following social risk and other factors on CollaboRATE scores using mixed effects logistic regression analysis: survey administration mode; patient age; patient gender; number of health conditions; patient's primary language; patient race; and percent of residents below the federal poverty line in the patient's home zip code. While collaboRATE is not currently specified to provide performance scores at the individual clinician level, clinicians were included as a random effect in this model to account for clustering of patients by clinician (Barr 2017).
The social risk variables in this Single-Group Primary Care Survey (described in sections 1.5 through 1.8 above) were drawn from the medical group's electronic medical record system and matched to patient survey responses via a unique identifier.
As the results of the mixed effects logistic regression analysis described above (results presented in 2b3.4a) showed that only patient age and survey administration mode had statistically significant associations with CollaboRATE scores, we selected only these two factors for inclusion in the risk adjustment model.

**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?**

In the analysis described above in section 2b3.3a, CollaboRATE performance scores increased slightly with patient age (OR 1.01 per year of age, 95% CI 1.01 to 1.02) and no other patient characteristics were associated with CollaboRATE performance scores (Barr 2017). Mode of survey administration was also significantly associated with CollaboRATE performance scores (patient portal mode: OR 0.60, 95% CI 0.45 to 0.80; automated phone/IVR mode: OR 0.45, 95% CI 0.34 to 0.59; SMS mode: OR 0.51, 95% CI 0.38 to 0.67).

**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.**

As the analysis outlined in section 2b3.3a showed that empirical associations between observed patient characteristics and the CollaboRATE performance score outcome were limited to patient age and mode of survey administration, we included only patient age and mode of survey administration in the statistical risk model. Adjusting for these characteristics will allow for fair comparison of CollaboRATE performance scores between providers at high and 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*)

We first estimated a logistic regression model incorporating as independent variables the two characteristics shown to have a statistically significant association with CollaboRATE performance scores, i.e. patient age and mode of survey administration. We used robust standard errors to account for the clustering of CollaboRATE outcome data by clinician. In a step-wise manner, we incorporated interaction terms to the extent that their odds ratios were statistically significant and that they improved the fit of the model. The final model was a logistic regression with robust standard errors and CollaboRATE performance score as the dependent variable with patient age and mode of survey administration as independent predictor variables.

*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*)**:**
The c-statistic (concordance statistic) for this statistical risk model is 0.6353 (95% CI 0.6178-0.6529). As the c-statistic's confidence interval does not include 0.5, we conclude that this model is superior to random chance in predicting CollaboRATE outcomes.
The pseudo R-squared value for this model is 0.0354, indicating that approximately 3.5% of variation in CollaboRATE scores is explained by the independent variables patient age, mode of survey administration, and interaction between age and survey mode. Given the role of clinician performance in measurement of shared decision-making, this low pseudo R2 value is expected.

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

The Hosmer-Lemeshow statistic for this statistical risk model is 773.46; p=0.1034. The null hypothesis in the Hosmer-Lemeshow test is that the model is correctly specified; therefore, our failure to reject the null hypothesis based on a p-value of 0.1034 indicates that the model is correctly specified.

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

N/A

**2b3.9. Results of Risk Stratification Analysis**:

N/A

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

The tests above demonstrate adequacy of this statistical risk model in controlling for differences in patient characteristics and survey administration mode. The c-statistic of 0.6353 is well above the threshold (0.05) at which the model does no better than random chance at predicting the outcome variable. The pseudo R2 value of 0.0354 suggests that there are other factors, prominently including individual clinician SDM performance, that explain variation in CollaboRATE scores.

The Hosmer-Lemeshow statistic of 773.46 (p=0.1034) indicates that the model is correctly specified. The null hypothesis in the Hosmer-Lemeshow test is that the model is correctly specified; therefore, our failure to reject the null hypothesis based on a p-value of 0.1034 indicates goodness of fit in the current model.

**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)*

In the Three Group Primary Care Survey study, mixed effects regression analysis adjusted for fixed effects (including patient age, patient gender, medical group, and mode of survey administration) across three clinical groups and also included individual clinician as a random effect to account for clustering of patient responses by clinician and estimate the extent to which CollaboRATE scores varied by individual clinician. We examined the random effect variance estimate to determine how much variation exists in individual clinician

performance when controlling for patient case mix. We conducted post-estimation z-tests to compare medical group performance based on the model controlling for patient case mix.

**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*)

We obtained a clinician random effect variance estimate of 0.146, which translates to a standard deviation of 0.382. With regard to fixed effects, we observed statistically significant odds ratios for site. Paired with a post-estimation z-test, these odds ratios demonstrated that Group 3 (OR 1.759, 95% CI 1.216-2.545) attained significantly higher CollaboRATE scores than the reference category Group 1 (OR 1.00) or Group 2 (OR 0.922, 95% CI 0.570-1.492; z=-2.71, 95% CI −1.114 to −0.178, p=.007) (Forcino 2018).

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The clinician random effect variance of 0.146 implies that the distribution of CollaboRATE scores varied substantially between clinicians. As the corresponding standard deviation of 0.382 translates to 0.5364 on the probability scale, a clinician whose scores fall one standard deviation above the mean clinician will have a 53.64% greater probability of obtaining a perfect CollaboRATE score from a randomly selected patient (Forcino 2018). Additionally, fixed effects for medical group paired with post-estimation hypothesis tests demonstrated that Group 3 (OR 1.759, 95% CI 1.216-2.545) attained significantly higher CollaboRATE scores than the reference category Group 1 (OR 1.00) or Group 2 (OR 0.922, 95% CI 0.570-1.492; z=-2.71, 95% CI −1.114 to −0.178, p=.007) (Forcino 2018).

---

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

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

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

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

N/A

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

N/A

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

N/A

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

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

In the Single Group Primary Care Survey, we conducted descriptive analysis of CollaboRATE data, including frequency calculations. We also compared the demographic characteristics of respondents to non-respondents for each mode. Pearson's χ2 tests and Student's t-tests were used for categorical and continuous variables, respectively. Logistic regression analysis was used to confirm the descriptive findings comparing respondents to non-respondents and to examine whether an interaction between age and number of comorbidities predicted response, where response was the binary outcome variable and independent variables included age, number of comorbidities, age multiplied by number of comorbidities, gender, age, and whether the visit was for a wellness check-up.

**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*)

In the Single Group Primary Care Survey study, descriptive analysis of CollaboRATE data found that more than 99% of respondents completed all three CollaboRATE items, resulting in very little missing data.

With regard to the CollaboRATE performance score and potential response bias, respondents in the Single Group Primary Care Survey study tended to be slightly older than non-respondents across all survey administration modes, and representative of the overall clinic population with regard to gender (Barr 2017).

**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*)

Respondents in the Single Group Primary Care Survey study were slightly older than non-respondents, and represented the overall clinic population with regard to gender. Respondents were slightly more likely than non-respondents to be seen for an annual wellness visit. Given these few demographic differences between respondents and non-respondents, and the very small magnitude of these differences, we expect that CollaboRATE performance scores are not significantly biased by non-response.

## 3. Feasibility

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

**3a. Byproduct of Care Processes**

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

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

Other

If other: Patient-reported CollaboRATE data may be collected by clinic staff at the point of care, through the clinic's electronic or telephone outreach to patients following their clinic visits, or by a third party.

**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**.

Patient/family reported information (may be electronic or paper)

**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).

CollaboRATE has been tested on a tablet, and using a web-based platform (online patient portal weblink sent by email). We have also successfully collected data using responses to text messages (SMS) on cellular telephones.

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

**Attachment:**

**3c. Data Collection Strategy**

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

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

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

We have learned that ensuring patient confidentiality at the time of survey completion allows for greater variation in the range of scores. Feasibility issues with regard to data collection are more prominent with paper surveys administered in clinic; automated patient surveys administered via text message or online patient portal require less staff time to administer and allow respondents more flexibility.

We recommend limiting data collection to 25 patients per unit of analysis to minimize response burden for patients. Clinicians whose performance is being measured have expressed interest in the measure, asking for details on how they can improve their performance; we recommend making these resources available to entities whose performance is being measured.

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

The measure is licensed under Creative Commons and freely available online at http://glynelwyn.com/collaborate-measure.html. There are no fees associated with use of the measure.

## 4. Usability and Use

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

**4a. Accountability and Transparency**

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

**4.1. Current and Planned Use**

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

| Specific Plan for Use | Current Use (for current use provide URL) |
|---|---|
| | Payment Program<br>Blue Shield of California<br>http://healthaffairs.org/blog/2016/03/28/an-innovative-patient-centered-total-joint-replacement-program/<br>Quality Improvement (external benchmarking to organizations)<br>Pacific Business Group on Health<br>http://www.pbgh.org/programs/21-the-patient-assessment-survey<br>Right for Me<br>http://www.rightforme.org/right-for-me.html<br>Evaluating CollaboRATE<br>Evaluating CollaboRATE, http://bmjopen.bmj.com/content/7/3/e014681<br>Quality Improvement (Internal to the specific organization)<br>United States Department of Veterans´ Affairs<br>https://www.va.gov/ |

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

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

Name of program and sponsor: Blue Shield of California
Purpose: Quality improvement
Geographic area and number and percentage of accountable entities and patients included: State of California; 100% of Blue Shield members requesting pre-authorization for total joint replacement or hysterectomy procedures.
Level of measurement: Other (Individual patient)
Setting: Outpatient Services
Name of program and sponsor: US Department of Veterans Affairs; Dr. Barbara Bokhour
Purpose: Research
Geographic area and number and percentage of accountable entities and patients included: VA Medical Center, Bedford, MA; 1019 outpatient respondents and 767 inpatient respondents.
Level of measurement: Clinician: Individual
Setting: Inpatient/Hospital; Outpatient Services
Name of program and sponsor: Pacific Business Group on Health
Purpose: Quality Improvement

Geographic area and number and percentage of accountable entities and patients included: State of California; 31,000 patient respondents who visited one of 150 participating provider groups, demonstrating variation in CollaboRATE performance scores across clinicians and provider groups, with provider group-level scores ranging from 40% to 68%.
Level of measurement: Clinician: Group/Practice
Setting: Outpatient Services
Name of program and sponsor: Right for Me, Dartmouth College and PCORI
Purpose: Research
Geographic area and number and percentage of accountable entities and patients included: Cluster randomized controlled trial including 16 clinics/provider groups in New England demonstrated variation in CollaboRATE performance scores across clinics.
Level of measurement: Clinician: Individual; Clinician: Group/Practice
Setting: Outpatient Services
Name of program and sponsor: Evaluating CollaboRATE, Dartmouth College and Moore Foundation
Purpose: Research
Geographic area and number and percentage of accountable entities and patients included: Three clinics/provider groups in New England and California, including more than 5,000 patient CollaboRATE responses, demonstrated variation in CollaboRATE performance scores across clinics. At one clinic with more than 4,000 patient responses, clinician-level scores changed with varying survey administration modes, though clinician rank order remained consistent across all survey administration modes.
Level of measurement: Clinician: Individual; Clinician: Group/Practice
Setting: Outpatient Services

**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. There are no policies in place to restrict access to performance results or impede implementation.

**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.**

Following 15 months of data collection in a primary care clinic, participating clinicians received single-page reports detailing their results and comparing their scores to those of their anonymous colleagues. These reports were accompanied by a seminar in which Glyn Elwyn presented shared decision-making theory and best practices.

**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.**

Score reports were provided to participating clinicians at the end of data collection. An educational seminar was scheduled during a clinic faculty meeting where Glyn Elwyn presented shared decision-making theory, best practices, and methods for improving CollaboRATE scores.

**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.**

Measured entities have provided feedback on implementing the collaboRATE tool into their practices based on diverse implementation strategies and experiences. The brief nature of the measure and its ability to be completed by most patients in under 30 seconds contributes to its ease of use. Paper-based survey implementation within the clinic environment was seen as a challenge due to the personnel and related resources required to administer the paper questionnaire to patients. This challenge was surmountable, particularly through automated questionnaire delivery via text messages (SMS), automated phone calls, and online patient portals (e.g. MyChart). Feedback related to automated questionnaire administration has been positive due to its limited impact on clinic workflows and few ongoing resource requirements after initial set-up.

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

N/A

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

N/A

**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.**

N/A

**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.**

Performance results can be compared across entities to illustrate variation in shared decision making across clinicians, practices, facilities, and health systems. Widespread data collection is currently underway by payers (e.g. Blue Shield of California) and entities such as the Pacific Business Group on Health (who have collected more than 30,000 CollaboRATE responses to date across multiple health care practices in California) to determine how shared decision making performance varies across providers and systems with a goal of improving shared decision making and resulting health care quality for individuals and populations.

**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 unintended negative consequences to individuals or populations were identified during testing. Clinical testing included 15 months of routine data collection in a primary care clinic, where all patients visiting participating clinicians were eligible to complete CollaboRATE. A total of 4421 CollaboRATE responses were collected during this testing period with no unintended negative consequences identified.

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

Feedback from patients completing the survey demonstrates that individuals value being asked specifically about shared decision making and being able to provide feedback on that topic (Elwyn 2013, doi: 10.1016/j.pec.2013.05.009).

Clinicians who participated in a 15-month CollaboRATE data collection period gathered for a focus group at the end of data collection, at which they expressed interest in the CollaboRATE measure and in their performance scores and how they might improve their shared decision-making performance.

## 5. Comparison to Related or Competing Measures

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

**5. Relation to Other NQF-endorsed Measures**

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

Yes

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

2962 : Shared Decision Making Process

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

**5a. Harmonization of Related Measures**

The measure specifications are harmonized with related measures;

**OR**

The differences in specifications are justified

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

**Are the measure specifications harmonized to the extent possible?**

Yes

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

The measure specifications are not completely harmonized as CollaboRATE's target population is more inclusive than measure 2962 of Shared Decision Making Process, which focuses on patients undergoing specific surgical procedures. Instead, CollaboRATE allows for assessment of shared decision-making performance relevant to any type of health care encounter, context, or setting.

**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_2_-_Testing_attachment_-_Medical_group_signal-to-noise_calculations_final_for_Jan_2019.xlsx

## Contact Information

**Co.1 Measure Steward (Intellectual Property Owner):** The Dartmouth Institute for Health Policy & Clinical Practice

**Co.2 Point of Contact:** Glyn, Elwyn, glynelwyn@gmail.com, 603-729-6694-

**Co.3 Measure Developer if different from Measure Steward:** The Dartmouth Institute for Health Policy & Clinical Practice

**Co.4 Point of Contact:** Glyn, Elwyn, glynelwyn@gmail.com, 603-729-6694-

## 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.**

Individuals involved in the development of the measure, including interviews with target end users and pilot testing, include: Glyn Elwyn, Paul Barr, Rachel Thompson, Stuart Grande, Thom Walsh, Elissa Ozanne, and Rachel Forcino.

**Measure Developer/Steward Updates and Ongoing Maintenance**

**Ad.2 Year the measure was first released:** 2013

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

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

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

**Ad.6 Copyright statement:** CollaboRATE is registered under a Creative Commons license and is freely available for use.

**Ad.7 Disclaimers:**

**Ad.8 Additional Information/Comments:**