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 #: 2523

Corresponding Measures:

De.2. Measure Title: Rheumatoid Arthritis: Assessment of Disease Activity

Co.1.1. Measure Steward: American College of Rheumatology

De.3. Brief Description of Measure: Percentage of patients 18 years and older with a diagnosis of rheumatoid arthritis and >=50% of total number of outpatient RA encounters in the measurement year with assessment of disease activity using a standardized measure.

1b.1. Developer Rationale: Disease activity is a key outcome in RA. American College of Rheumatology (ACR) guidelines recommend routine disease activity measurement in clinical practice to target low disease activity or remission in all patients. Clinical trials indicate that using validated assessments to set treatment goals and target therapy results in improved patient outcomes, including better functional and radiographic outcomes.

S.4. Numerator Statement: # of patients with >=50% of total number of outpatient RA encounters in the measurement year with assessment of disease activity using a standardized measure.

S.6. Denominator Statement: Patients 18 years and older with a diagnosis of rheumatoid arthritis seen for two or more face-to-face encounters for RA with the same clinician during the measurement period.

S.8. Denominator Exclusions: N/A

De.1. Measure Type: Process

S.17. Data Source: Electronic Health Records, Registry Data

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

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

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

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

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

The developer provides the following evidence for this measure:

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

Evidence Summary

- Brief background: this measure evaluates patients 18+ years with rheumatoid arthritis and >=50% of outpatient RA encounters with assessment of disease activity.
- Disease activity (how patients with RA are responding to treatment or whether they are reaching treatment goals) is a key outcome for patients with RA.
- Developer provided a logic model that outlines the relationship between the measurement of RA activity, review of the assessment, identifying increased activity, modification of therapy, and the achievement of positive RA outcomes.
- Assessment by a validated instrument is a 2015 clinical practice guideline recommendation supported by systematic reviews of randomized controlled trials and individual cohort studies.

Questions for the Committee

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

Guidance from the Evidence Algorithm

Process measure based on a systematic review/grading of evidence (box 3) YES -> QQC based on SR (box 4) YES -> SR concludes QQC MODERATE (box 5b) -> MODERATE

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

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

Maintenance measures – increased emphasis on gap and variation

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

- Note: developer did not perform separate analysis according to their specifications, which are for both individual and group/practice level clinicians.
- Most recent performance data from 2017 has a mean of 43.91%, and performance ranging from 0% (first and second deciles) to 100% (tenth decile).

- 2017 data indicates a drop in performance; first reported quarter was 2014 Q3 with 44.6% and 2016 Q2 reports 62.97%.

Disparities

- Disparities data is not routinely collected or available in the RISE registry, however, the developer states optimal clinical performance should be 100%.
- Observational studies of patients with RA suggest significant disparities in disease activity and clinical outcomes across racial and ethnic groups; data from one large registry suggests differences in mean disease activity across racial and ethnic groups, with African-Americans being less likely to achieve clinical remission and having higher disease activity overall.

Questions for the Committee:

- Are there concerns with the drop in performance in the most recent measurement year?
- Is there a gap in care that warrants a national performance measure?
- Are you aware of evidence that disparities exist in this area of healthcare?

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

RATIONALE:

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

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

- Evidence shows a high correlation with following the process and improved care, however, no evidence seems available to contradict the basis for this measure.
- Evidence supports this measure being important for quality of care, but most EHRs do not allow it to be pulled as structured data. This measure would be hard to implement in many practices.
- assessment of disease activity is supported by guideline recommendations based upon RCT, systematic reviews and cohort studies
- Evidence applies directly to the outcome being measured.

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?

- The performance gap is significant enough to warrant this measure. Some data on disparities in outcome were described, but developer does not recognize any social risk factors that would suggest any disparities that would be appropirate for the population subgroups.
- could not be assessed in the RISE population since demographics not sufficiently collected.
- significant opportunity for improvement and some evidence from a registry for racial disparities on response to treatment;
- Althought clinically used to determine response to treatment and decision to change therapy, this measure has consistently performed poorly when measured as indicated by RISE data. Current performance data do not reveal health disparities as social risks data is not collected RISE registry.

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 Staff

Evaluation of Reliability and Validity:

- Note that the measure developer has pooled the data for individual and practice level performance to perform their analyses, and therefore the measure has not been tested to specifications. Measures must be tested to specifications, meaning separate reliability analyses conducted for each level of analysis. In this case, separate analyses for clinician: individual and clinician: group/practice.
- Measure score reliability assessed using signal-to-noise, with a mean reliability score 0.97, ranging from 0.48-1.00.
- For validity, the developer assessed critical data element validity using data abstracted from randomly sampled patient records, which were used to calculate parallel forms reliability for the measure; and inter-rater agreement using a kappa coefficient, to assess whether EHR specifications and data exported electronically from the EHR were valid when compared to a front-end chart abstraction of the entire EHR by trained reviewer. Face validity testing was also conducted.

Questions for the Committee regarding reliability:

- Should the developer indicate a minimum number of cases needed, to ensure reliability? Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

Committee should discuss the implications of the reliability testing and the need to perform analyses according to specifications.***Questions for the Committee regarding validity:***

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

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

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

Rationale

- **The measure developer has pooled the data for individual and practice level performance to perform their analyses, and therefore the measure has not been tested to specifications.**
- **Measures must be tested to specifications, meaning separate reliability analyses conducted for each level of analysis.**
- **In this case, separate analyses for clinician: individual and clinician: group/practice. The measure has therefore been scored as insufficient.**

Evaluation A: Scientific Acceptability

Measure Number: 2523

Measure Title: Rheumatoid Arthritis: Assessment of Disease Activity

Type of measure:

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

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

Data Source:

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

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

☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

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

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

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

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

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

2. **Briefly summarize any concerns about the measure specifications.**
   - Should the developer indicate a minimum number of cases needed, to ensure reliability?

RELIABILITY: TESTING

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

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

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

   Submission document: Testing attachment, section 2a2.2

   - The measure developer has pooled the data for individual and practice level performance to perform their analyses, and therefore the measure has not been tested to specifications.
   - Measures must be tested to specifications, meaning separate reliability analyses conducted for each level of analysis.
   - In this case, separate analyses for clinician: individual and clinician: group/practice. The measure has therefore been scored as insufficient.
   - Measure score reliability was assessed using a signal to noise analysis.
   - Data for reliability testing was collected from outpatient rheumatology clinics that participate in the ACR's Rheumatology Informatics System for Effectiveness (RISE) reg the
   - Practices included in the signal-to-noise analysis were limited to those that were evaluated on measure performance from January 2017 through December 2017, which totaled 107. Of these 107, 27 (25%) practices were individual providers, while the other 80 (75%) were group practices or health systems.
7. **Assess the results of reliability testing**

   - Mean reliability score 0.97, ranging from 0.48-1.00. Developer states that a few extreme outliers with poor reliability can likely be attributed to low case volume.
   - The results demonstrate strong reliability.

   Submission document: Testing attachment, section 2a2.3
8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)
9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (data element testing was not performed)
10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

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

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

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

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

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**
    - The measure developer has pooled the data for individual and practice level performance to perform their analyses, and therefore the measure has not been tested to specifications.
    - Measures must be tested to specifications, meaning separate reliability analyses conducted for each level of analysis.
    - In this case, separate analyses for clinician: individual and clinician: group/practice. The measure has therefore been scored as insufficient.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.
    - No exclusions

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

    Submission document: Testing attachment, section 2b4.
    - Developer states that "differences between providers in use of validated disease activity assessments in RA reflects a meaningful gap in quality, based on qualitative feedback from clinicians and statistical analysis of the data." However, developer also notes that some differences in performance across sites is attributable to whether or not clinicians are entering structured data as part of the current workflow.
    - 2017 review found average performance was 43.91%, ranging from less than 1% to 100% (100% is optimal clinical performance)

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

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

    Submission document: Testing attachment, section 2b6.
    - Developer states there is no missing data in the registry. Developer states that if a data element is missing, indicates the provider did not perform expected action, not that the data itself is missing.
    - However, developer notes that this measure requires the use of a validated disease assessment instrument, some of which may require a patient-reported component, and therefore patient non-response may lead to missing data and inability to capture a disease activity score.
    - The developer stated they found missing data based on patient non-response to be a rare occurrence. There are no procedures for handling missing data.

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

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

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

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

16d.**Risk adjustment summary:**

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

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

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

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

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

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

N/A

VALIDITY: TESTING

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

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

☒ **Face validity**

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

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

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

Submission document: Testing attachment, section 2b2.2

- Developer assessed critical data element validity using data abstracted from randomly sampled patient records, which were used to calculate parallel forms reliability for the measure.
- Developer assessed inter-rater agreement using a kappa coefficient, to assess whether EHR specifications and data exported electronically from the EHR were valid when compared to a front-end chart abstraction of the entire EHR by trained reviewers. Each record underwent front-end review by two separate reviewers, and conflicts in this front-end data were adjudicated by the project lead investigator
- Additional validation performed during RISE registry onboarding/yearly audit process.
  - Developer notes this is functionally a registry measure, that cannot be reproduced, but can be assessed through iterative work between practices, registry tech vendor, and data analytic centers.
  - RISE dashboard allows providers to evaluate against registry average.
- Yearly audits conducted to verify accuracy of the patient data extracted from the EHR systems of a random sample of participating practices
- Face validity testing during measure development process.

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

Submission document: Testing attachment, section 2b2.3

- Original Kappa scores (from testing eMeasure): Kappa *Overall*, Range, % Agreement: **0.81** (0.73 to 0.89), 0.91, which is considered very reliable.
    - Kappa, Range, % Agreement Denominator: 0 (0, .97) 98.2%
        - Developer notes: Because instances of agreement dominated, the denominator Kappa was zero. The instance of 0 for the denominator is an example of the limitation of the Kappa statistic. A kappa of zero can be obtained even though agreement is very high due to one classification category dominating.
    - Kappa, Range, % Agreement Numerator: **0.84**, 0.77 to 0.91, 92.2%
    - Kappa, Range, % Agreement Exceptions: **1.00** (1.0 to 1.0), 100% (100% agreement that there are no exceptions)
- 2018 registry manual audit of 2017 data found 97.99% success rate (correct responses).
- Median face validity score was 9 using RAND/UCLA rating scale; median feasibility score was 7. Of 14 raters, 11 had a validity score greater than or equal to 7. Public comments and input from committees and the Board of ACR was also collected.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

    ☒ **Yes**

    ☐ **No**

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

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

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

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

    ☐ **Low** (NOTE: Should rate LOW if you believe that there <u>are</u> threats to validity and/or relevant threats to validity were <u>not assessed OR</u> 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 <u>is required</u>; if not conducted, should rate as INSUFFICIENT.)

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

ADDITIONAL RECOMMENDATIONS

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

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

2a1. Reliability-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?

- There are concerns about the reliability and may be insufficient in that the methodology doesn't meet NQF guidelines.
- Majority of practices do not collect any of the recommended measures as structured data. Some EHR will calculate some of the scores, few will do all of the scores, and for the scores that require labs to be incorporated or where the calculation is proprietary the score cannot be incorporated into the EHR.
- pooling of data (individual and group) does not allow for meeting specification requirements

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

- It may need a different level of testing to meet the specifications.
- measures are reliable. I am concerned simply about reporting
- see 2a1
- No

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

- The validity testing seems to be adequate.
- measures are reliable. I am concerned simply about reporting
- none
- No

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

- Missing data was reviewed but seemed to be very negligible and would not be a threat to validity.
- Missing data will be a huge problem particularly for the scores that are composites of physician evaluation and labs. Multiple choices of performance score -- will they be considered to be equal as performance measures. Missing data will be a problem
- as the standardized measurement tool may be PRO, patient non-compliance may be issue however, developer states this is rare occurrence
- 2b4 - data from RISE registry given important information about practice differnces among providers within the same institution along with a consistent improvement in performance with routine feedback. 2b5 - results are comparable. 2b6 - No missing data

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

- Risk adjustment doesn't apply. However, in future there may be a need to collect data about social issues, such as migrant status or race.
- case mix adjustment needs to be done. Also need to consider access to drugs -- not all insurances approve escalation of therapy.
- no exclusions, no risk adjustment

- 2b2 - no exclusions. 2b3 - No risk adjustment needed

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.

- Data elements generated during routine care; all data elements in defined fields in electronic sources
- This measure was approved for trial use as an eCQM. However, the developer reported that current HQMF specifications were insufficient to capture all the data elements required for measurement.
- The developer also noted that ACR has practices participating in the ACR´s RISE registry using more than 30 different electronic health record vendors. Based on member input, ACR made a conscious decision to avoid forcing individual practitioners to change their workflow and documentation to satisfy requirements for HQMF specifications.
- ACR also noted that the majority of RISE participants are solo or small practices and unaffiliated with an academic or other institution, and few have IT services sufficient to support modifications to their electronic health records to meet eCQM standards.
- For these reasons, ACR decided to change this from an eMeasure to a standard quality measure. This submission is a registry measure using EHR data.
- The developer states they "will continue to monitor developments in coding and HQMF specifications to determine if the updates would provide the necessary flexibility to make this measure an eCQM."
- Developer notes that measurement of RA disease activity using a standardized, validated instrument requires significant changes to current clinical workflow for many practices. In addition, different tools have varying levels of feasibility for practices, depending on resources available (labs, support staff).
- No fees or licensing required

Questions for the Committee:

- Is the data collection strategy ready to be put into operational use?

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

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?

- By using the RISE registry, the measurement data collection strategy seems feasible.
- Data collection will be a problem. Not all of these measures are uniformly collected, not all EHR support this data collection Often it is in free text in chart notes, not in structured data.
- agree with moderate rating, with caveat that use of standardized measurement tool may require change in workflow of the clinician/practice
- Since various validated disease activity measures exist, their implementation is feasible in clinical practice

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

- Used in MIPS; approximately 3,550 providers eligible
- RISE registry – ACR: 937 providers, 1.78 million patients. Internal QI and external benchmarking.
- Previously used in PQRS.

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

- Providers have access to results via registry (updated monthly), can contact ACR or registry vendor staff with issues and questions
- This measure is new to MIPS; developer will be able to get feedback from MIPS users starting in 2020.

Additional Feedback: N/A

Questions for the Committee:

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

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

RATIONALE:

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

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

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

Improvement results

- Performance rates are decreasing over time (worse results in Q4 2017 than Q2 2016). Developer suggests early positive results may be skewed by early adopter phenomenon and later results are more reflective of a more generalizable group of US rheumatology practices as participation rates have more than doubled.
- The developer notes that this version of the measure replaces a less stringent one (reporting once a year) and is new to MIPS for 2019.

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

- Developer found key data elements were in non-standardized formats, leading to challenges in pulling HQMF-formatted data.
- No negative/unintended impact on patients found.
- Providers gave positive feedback, noting better understanding of practice variations and better identification of higher-risk patients.

Potential harms

- None found

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**

RATIONALE:

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

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

- Users have been given ample opportunity to provide feedback.
- This is being done through RISE but the RISE registry represents only a small proportion of rheumatologists in the USA.
- agree with Pass - currently collecting and sharing data and plans to expand this via MIPS
- Yes.

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.

- With some issues over time, there will need to be work on the improvement. However, there do not seem to be harms or unintended consequences to the measure.
- No harms, but often there are many factors (including patient compliance and access to care) that impact therapy choices.
- agree with Moderate rating - developer explains that worsening results may be consequential to "early adopter" phenomenon
- Using valiadate disease activity measurement improves patient outcomes as evidenced in the literature. The performace results can be used as an impetus to encourage physicians to document disease activity in a formal manner during their visit rather than in free text. No unintended consequences foreseen.

Criterion 5: Related and Competing Measures

Related or competing measures
N/A
Harmonization

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

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

- There are no related or competing measures.
- None to my knowledge.
- None
- No

Public and Member Comments

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

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

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

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

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

Disease_Activity_Measure_Evidence_Form_Final-635294352077200854.docx,RA_DAS_evidence_form_2019_FINAL.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

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

No

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 2523

Measure Title: Rheumatoid Arthritis: Assessment of Disease Activity

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

Date of Submission: 4/1/2019

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: Rheumatoid Arthritis: Assessment of Disease Activity (collection of outcome score)

☐ 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.

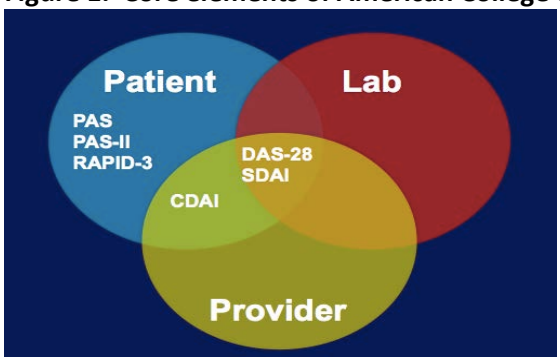
The proposed measure is a *process* measure that requires collection of a key *health outcome* using a standardized score. Collecting this outcome measure in routine clinical care is supported by American College of Rheumatology (ACR) guidelines (*Singh J et al.* 2015 American College of Rheumatology

Guideline for the Treatment of Rheumatoid Arthritis*. Arthritis Rheumatol. 2016 Dec;68(1):1-26*).

The ACR undertook an extensive multi-year project, involving systematic literature reviews, expert consensus ratings, and national surveys to reach consensus on which RA disease activity measures are valid, reliable, and responsive, and feasible to implement in routine clinical practice (*Anderson J et al., Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken). 2012 May;64(5):640-7*). This manuscript is included as a supplemental Appendix.

The ACR endorsed 6 RA disease activity measurement tools, which include overlapping core elements (Figure 1). All include a patient-reported component (PRO). No measure is currently a gold standard; there is good scientific evidence supporting each endorsed measure. Therefore, clinicians can select from a range of valid options appropriate to their practice settings and available resources. This novel approach to measurement has been extensively validated in RA over a period of several decades (*Anderson J et al., Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken). 2012 May;64(5):640-7)*.

Figure 1. Core elements of American College of Rheumatology's endorsed rheumatoid arthritis



The 6 proposed outcome measures have cut points for low, moderate and high disease activity as well as disease remission to facilitate clinical decision-making. See Table 1.

Table 1. Disease activity cut points for American College of Rheumatology–recommended disease activity measures.

	Range	Remission	Low	Moderate	High
DAS28 (ESR or CRP)	0-9.4	< 2.6	≥ 2.6 - < 3.2	≥ 3.2 - ≤ 5.1	> 5.1
CDAI	0-76	≤ 2.8	> 2.8 - 10.0	> 10.0 - 22.0	> 22.0
SDAI	0-86	0.0 - 3.3	3.4 - 11.0	11.1 - 26.0	26.1 - 86.0
RAPID-3	0-10	0 - 1.0	> 1.0 - 2.0	> 2.0 - 4.0	> 4.0 - 10
PAS	0-10	0.00 - 0.25	0.26 - 3.70	3.71 - 7.99	8.00 - 10.00
PASII	0-10	0.00 - 0.25	0.26 - 3.70	3.71 - 7.99	8.00 - 10.00

In order to assess how patients with rheumatoid arthritis (RA) are responding to therapy or whether they are reaching treatment goals, RA disease activity should be assessed using a validated instrument.

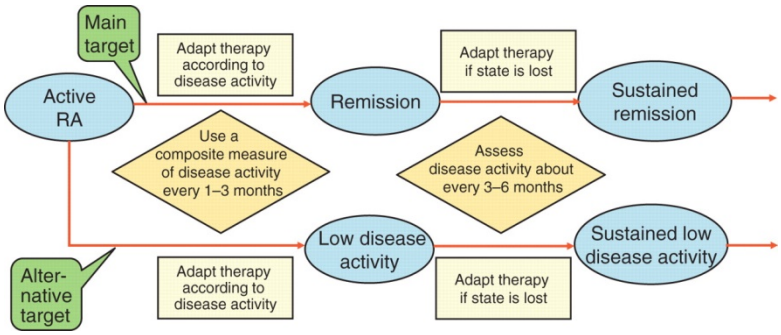
Step 1: Measure disease activity using validated instrument

Step 2: Review disease activity assessment with patient during office visit: is the patient in remission, low, medium (moderate) or high disease activity?

Step 3: If the patient has moderate or high disease activity, consider treatment modification with goal of remission/ low disease activity.

Step 4: At next office visit or 3-6 months after initiation/ change in medication, repeat Steps 1-3 until patient is in remission/ low disease activity or until patient is satisfied with their functional status (patient-reported outcome measure, a separate quality measure).

Figure 2. Algorithm for using standardized disease activity measures to target therapy in rheumatoid arthritis. *From Smolen et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis 2010;69:631-637 Algorithm for treating rheumatoid arthritis (RA) to target based on the recommendations.*



Standard collection of disease activity outcomes in RA to facilitate a "treat to target" approach, where the target is disease remission or low disease activity, has been shown to improve clinical and radiographic outcomes (*Schipper LG et al. A tight control treatment strategy aiming for remission in early rheumatoid arthritis is more effective than usual care treatment in daily clinical practice: a study of two cohorts in the Dutch Rheumatoid Arthritis Monitoring registry. Ann Rheum Dis. 2012 Jun;71(6):845-50; Smolen JS et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis. 2010 Apr;69(4):631-7; Grigor C et al. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomized*

controlled trial. Lancet 2004;364:263–9.).

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.)

While not all recommended disease activity measure are derived from patient report, it is important to note that patients participated in creating the updated 2015 ACR Guideline for the Treatment of Rheumatoid Arthritis and supported the key principle of collecting disease activity.

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

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

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

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

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• 2015 American College of Rheumatology • Guideline for the Treatment of Rheumatoid Arthritis • *Singh J et al.* • *2016 Dec* • *Arthritis Rheumatol.;68(1):1-26* • https://onlinelibrary.wiley.com/doi/epdf/10.1002/art.39480
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.	Disease activity measurement using an ACR-recommended measure should be performed in a majority of encounters for RA patients (16).† † Any of the ACR recommended disease activity measures may be chosen, as described in Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, et al. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken) 2012; 64:640–7.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Collection of disease activity is a fundamental principle underlying all remaining guidelines (evidence for subsequent guidelines ranges from low to high).
Provide all other grades and definitions from the evidence grading system	We developed this guideline following the recently revised ACR guideline development process (http://www.rheumatology.org/Practice-Quality/Clinical-Support/Clinical-Practice-Guidelines). This process includes the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) methodology (available at http://www.gradeworkinggroup.org) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013;66:719–25. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation: determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726–35.
Grade assigned to the **recommendation** with definition of the grade	N/A
Provide all other grades and definitions from the recommendation grading system	N/A

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	N/A
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2013 update • Smolen JS et al • Mar 2014 • Ann Rheum Dis; 73(3): 492–509 • https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3933074/
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.	Treatment should be aimed at reaching a target of remission or low disease activity in every patient. Monitoring should be frequent in active disease (every 1–3 months); if there is no improvement by at most 3 months after the start of treatment or the target has not been reached by 6 months, therapy should be adjusted.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Levels of evidence (LoE), grades of recommendations (GoR), strength of recommendation (SoR; = level of agreement), and % of votes for the respective items as worded, based on the recommendations of the Oxford Centre for Evidence-Based Medicine Treatment should be aimed at reaching a target of remission or low disease activity in every patient. LoE:1a; GoR: A; SoR: 9.6±0.7; %: 100. Monitoring should be frequent in active disease (every 1–3 months); if there is no improvement by at most 3 months after the start of treatment or the target has not been reached by 6 months, therapy should be adjusted. LoE:2b; GoR: B; SoR: 9.5±1.0; %: 100.
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	See above

<table>
<tr><td>Provide all other grades and definitions from the recommendation grading system</td><td>Figure 3. Recommendations of the Oxford Centre for Evidence-Based Medicine for levels of evidence (LoE) and grades of recommendations (GoR)
<table>
<tr><th>Level</th><th>Therapy/Prevention, Aetiology/Harm</th></tr>
<tr><td>1a</td><td>Systematic review (with homogeneity) of RCTs</td></tr>
<tr><td>1b</td><td>Individual RCT (with narrow Confidence Interval)</td></tr>
<tr><td>1c</td><td>All or none (ie all patients died before the Rx became available, but some now survive on it; or when some patients died before the Rx became available, but none now die on it)</td></tr>
<tr><td>2a</td><td>Systematic review (with homogeneity) of cohort studies</td></tr>
<tr><td>2b</td><td>Individual cohort study (including low quality RCT; e.g., <80% follow-up)</td></tr>
<tr><td>2c</td><td>“Outcomes” Research or ecologic studies (studies of group ch^{ics})</td></tr>
<tr><td>3a</td><td>Systematic review (with homogeneity) of case-control studies</td></tr>
<tr><td>3b</td><td>Individual Case-Control Study</td></tr>
<tr><td>4</td><td>Case-series (and poor quality cohort and case-control studies)</td></tr>
<tr><td>5</td><td>Expert opinion or based on physiology, bench research or “first principles”</td></tr>
</table>
Oxford Centre for Evidence Based Medicine</td></tr>
<tr><td>Body of evidence:
• Quantity – how many studies?
• Quality – what type of studies?</td><td>Not reported</td></tr>
<tr><td>Estimates of benefit and consistency across studies</td><td>N/A</td></tr>
<tr><td>What harms were identified?</td><td>N/A</td></tr>
<tr><td>Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?</td><td>N/A</td></tr>
</table>

_____

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

1b. Performance Gap

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

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

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

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

Disease activity is a key outcome in RA. American College of Rheumatology (ACR) guidelines recommend routine disease activity measurement in clinical practice to target low disease activity or remission in all patients. Clinical trials indicate that using validated assessments to set treatment goals and target therapy results in improved patient outcomes, including better functional and radiographic outcomes.

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

Performance over time

Dates: July 1, 2014 through June 30, 2016

Practices: 44

Providers: 223

2014 Q3: 44.6%

2014 Q4: 42.42%

2015 Q1: 53.05%

2015 Q2: 54.96%

2015 Q3: 55.65%

2015 Q4: 58.63%

2016 Q1: 61.07%

2016 Q2: 62.97%

Most recent performance

Dates: January 1, 2017 through December 31, 2017

Practices: 107

Setting: 73% group, 25% solo practitioner, 2% health system

Patients: 94,872

Mean: 43.91%

Standard Deviation: 37.46%

Min: 0.00%

Max: 100.00%

Interquartile Range: 80.08%

Deciles

10%: 0.00%

20%: 0.00%

30%: 4.98%

40%: 26.78%

50%: 42.96%

60%: 55.57%

70%: 73.69%

80%: 84.41%

90%: 97.09%

100%: 100.00%

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Relevant disparities data are not routinely and uniformly collected on all patients within the RISE registry.

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

This measure is not risk-adjusted and the RISE registry has limited data on social risk factors. Furthermore, optimal clinical performance for this measure should be 100%, regardless of social risk, as this measure reflects the minimum performance standard. Nevertheless, as part of RISE's ongoing efforts to expand and improve, the American College of Rheumatology is exploring ways to obtain better social risk data to appropriately monitor performance disparities going forward. However, observational studies of patients with RA suggest significant disparities in disease activity and clinical outcomes across racial and ethnic groups. For example, data from a large US registry using the Clinical Disease Activity Index (CDAI), one of the recommended measures for disease activity assessment, found important differences in mean disease activity level across racial and ethnic groups, with African-Americans being less likely to achieve clinical remission and having higher disease activity overall. Standardized collection of disease activity assessments such as CDAI therefore has significant potential to unveil such differences and provide critical data for reducing disparities in RA outcomes.

Greenberg JD1, Spruill TM, Shan Y, Reed G, Kremer JM, Potter J, Yazici Y, Ogedegbe G, Harrold LR. Racial and ethnic disparities in disease activity in patients with rheumatoid arthritis. Am J Med. 2013 Dec;126(12):1089-98

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):*

Musculoskeletal : Rheumatoid Arthritis

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

Health and Functional Status : Change

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

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.)*

https://www.rheumatology.org/Portals/0/Files/RA-Disease-Activity-Measure.pdf

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** Disease_Activity_Updated_Value_Sets_2018-03-30.xls

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

Attachment **Attachment:**
RA_Disease_Activity_Measures_ACR_Recommendations_for_Use_in_Clinical_Practice_Paper.pdf

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

Clinician

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

Yes

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

Current HQMF specifications were insufficient to capture all the data elements required for measurement. Also, we have practices participating in the ACR´s RISE registry using more than 30 different electronic health record vendors. Based on member input, ACR made a conscious decision to provide the most flexible route to electronic health record data-based measurement and avoid forcing individual practitioners to change their workflow and documentation to satisfy requirements for HQMF specifications. Finally, as the majority of RISE participants are solo or small practices and unaffiliated with an academic or other institution, few have IT

services sufficient to support modifications to their electronic health records to meet eCQM standards. For these reasons, we decided to change this from an eMeasure to a standard quality measure.

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

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

# of patients with >=50% of total number of outpatient RA encounters in the measurement year with assessment of disease activity using a standardized measure.

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).

For purposes of this measure, "Rheumatoid Arthritis Disease Activity Measurement Tools" include the following instruments:

- -Clinical Disease Activity Index (CDAI)
- -Disease Activity Score with 28-joint counts (erythrocyte sedimentation rate or C-reactive protein) (DAS-28)
- -Patient Activity Scale (PAS)
- -Patient Activity Score-II (PAS-II)
- -Routine Assessment of Patient Index Data with 3 measures (RAPID 3)
- -Simplified Disease Activity Index (SDAI)

A result of any kind qualifies for meeting numerator performance.

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

Patients 18 years and older with a diagnosis of rheumatoid arthritis seen for two or more face-to-face encounters for RA with the same clinician during the measurement period.

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

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

One of the requirements for a patient to be included in the Initial Patient Population is that the patient has a minimum of 2 RA encounters with the same provider, all occurring during the measurement period.

If the patient qualifies for the Initial Patient Population, then every encounter for RA should be evaluated to determine whether disease activity using a standardized measurement tool was assessed. The logic represented in this measure will determine if the patient had a disease activity assessment performed at each visit during the measurement period (ie, Occurrence A of Encounter, Performed). The measure requires all of the eligible encounters to be analyzed in order to determine if the patient's disease activity was assessed at >=50% of encounters for RA. Once it has been determined if the patient meets >=50% threshold, all patient data across a single physician should be aggregated to determine the performance rate.

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

N/A

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.)*

N/A

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

N/A

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

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

Cases Meeting the Target Process / Target Population

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

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

A random sample is obtained by assigning each patient a sequential number and then using a random number generator to select patients.

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

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

N/A

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

If other, please describe in S.18.

Electronic Health Records, Registry Data

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

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

Data source 1: electronic health records

Instrument: RA Measure Testing Data Collection Form

Data source 2: Rheumatology Informatics System for Effectiveness (RISE) Registry

Data collection: passive abstraction from EHR

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 in attached appendix at A.1

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

Clinician : Group/Practice, Clinician : Individual

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

Outpatient Services

If other:

S.22. <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.)*

2. Validity – See attached Measure Testing Submission Form

Disease_Activity_Measure_Testing_Form_Final.docx,RA_DAS_measure_testing_form_January_2019_FINAL_4.3.2019_Update-636912728895407605.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.

Yes

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.

Yes

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.

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 2523
Measure Title: Rheumatoid Arthritis: Assessment of Disease Activity
Date of Submission: 1/7/2019

Type of Measure:

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

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

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

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

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

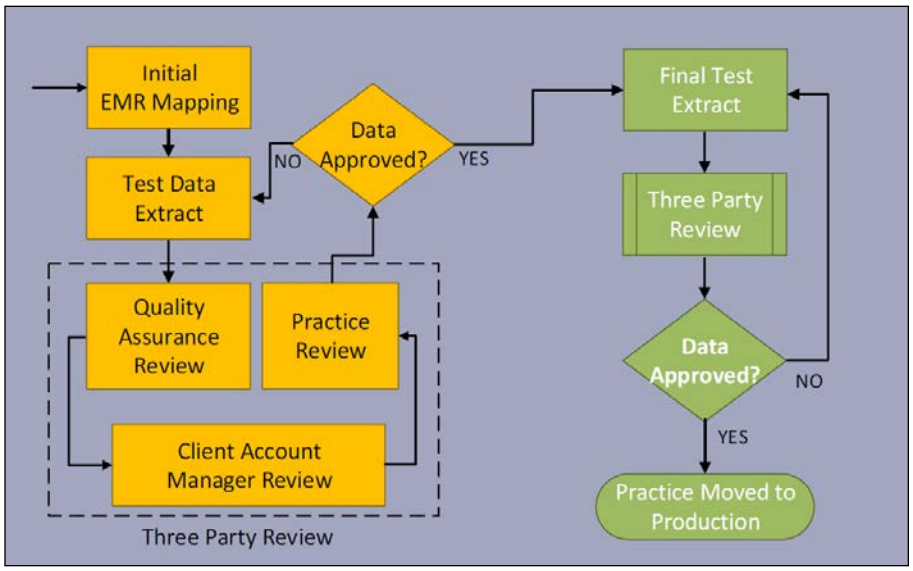
1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS,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

Registry data used for the most recent testing of this measure was collected through the ACR's Rheumatology Informatics System for Effectiveness (RISE) registry. RISE is a Qualified Clinical Data Registry (QCDR) that has been in operation since 2014. It was developed to serve as a tool for improving quality of care in rheumatology practices and a mechanism for providers to complete various federal reporting requirements for Medicare reimbursements. As of September 30, 2018, 218 practices across the United States with a total of nearly 1.5 million patients were fully connected to the RISE registry.

RISE uses proprietary computer programming to extract patient data from the EHR systems of participating providers. The data is then aggregated and used to calculate performance on a number of quality measures, including this measure. Practices that participate in RISE must complete an extensive data validation process, as seen in Figure 1, in order to be considered fully connected. During this process, practices work closely with RISE registry technical experts to gather the necessary information on the practice and identify where and how patient information, such as outcome measures, medications, laboratory results, diagnoses, etc., is stored in

the provider's EHR. After the initial mapping to the various EHR fields is complete, the RISE team works with the practice to systematically extract and review test data via the RISE dashboard. The extracted data is used to calculate performance on each quality measure in RISE. The practice and registry technical experts then review the measure performance by drilling down into the patients included in and excluded from each step of the measure and the specific patient data used in the measure calculations. This allows the practices to confirm that each part of the measure calculation (denominator, numerator, exclusions and exceptions) does not include false negatives or positives and uses only accurate information. If any inaccuracies are discovered, the data extraction and mapping are refined and the review process begins again. This continues until the practice and the RISE team can validate that all the measure scores and patient data used to calculate the performance are accurate.

Figure 1. Custom mapping and data validation for RISE registry participants



Once practices are fully connected, they continue to monitor their data accuracy through the analytic dashboard. Additionally, a limited data set extracted from the registry data is shared with a third-party center for wider analytic purposes. This data analytic center is a highly regarded academic center experienced in working with EHR data. The center performs a variety of additional accuracy and validation checks on the limited data set.

For each measure incorporated into the RISE registry, the various data elements identified in the value set (including ICD-10, LOINC and CPT codes) and measure specifications are used to build a comprehensive data dictionary in order to identify the various data elements across the different EHRs at each practice. The data dictionary is then used as the basis for the XML programming code that runs against the registry data to calculate measure performance. The flowchart of the programming for the Rheumatoid Arthritis: Assessment of Disease Activity measure can be seen in Figures 2a and 2b.

Figure 2a. Flowchart of calculation for the Rheumatoid Arthritis: Assessment of Disease Activity measure

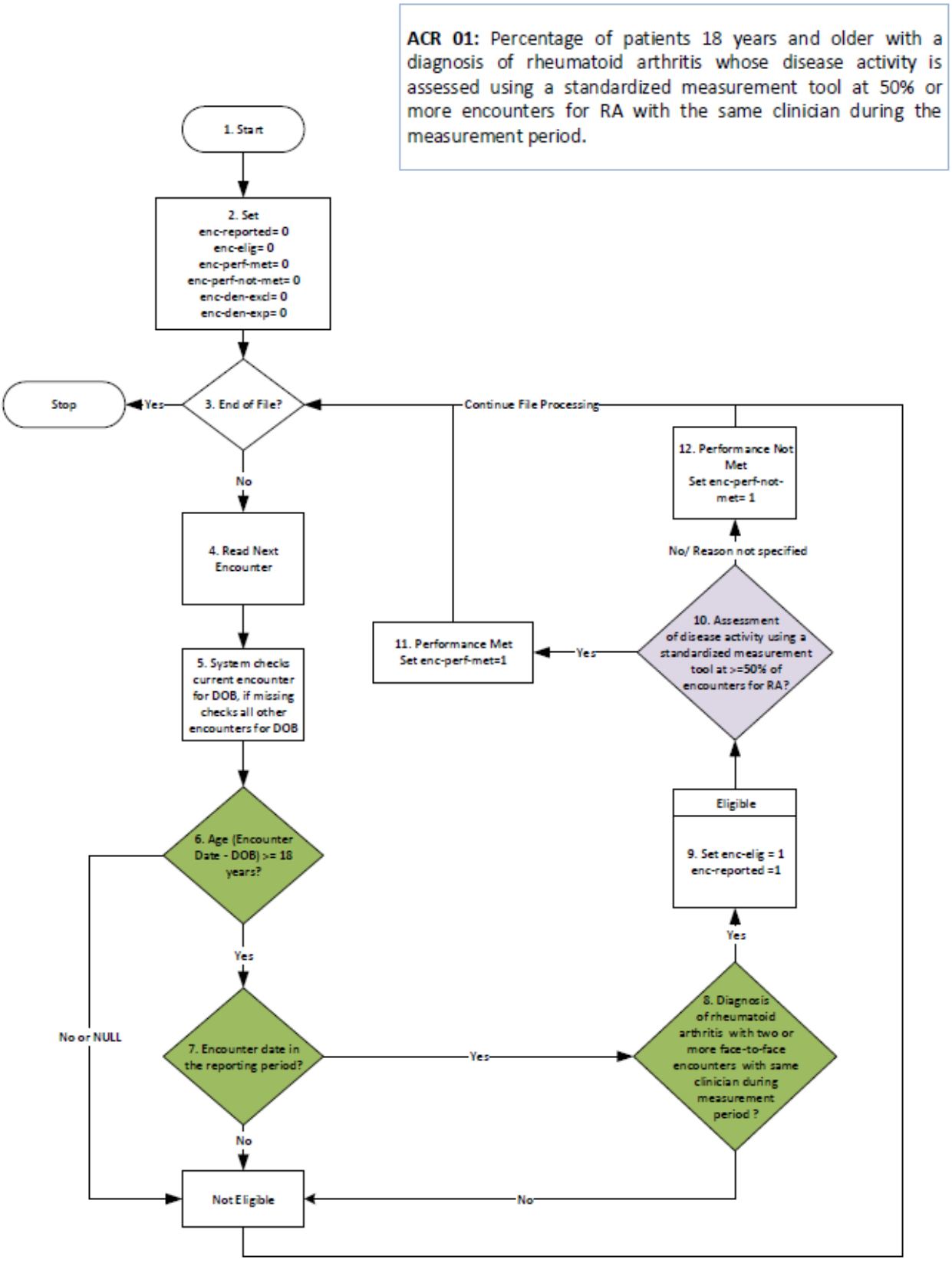


Figure 2b. Supplement to flowchart of calculation for the Rheumatoid Arthritis: Assessment of Disease Activity measure

Data Dictionary References

Denominator

Element ID	Element Name
1510	Encounter Date
2050	Date of Birth
4115	Encounter ACR 01, 02, 03, 04, 06, 07, 08
3220	Diagnosis of Active Rheumatoid Arthritis (RA)
3222	Date of diagnosis of Rheumatoid Arthritis (RA)

Numerator

Element ID	Element Name
3245	Disease activity assessed using a standardized measurement tool for RA- 1- Low
3245	Disease activity assessed using a standardized measurement tool for RA- 2- Moderate
3245	Disease activity assessed using a standardized measurement tool for RA- 3- High
3245	Disease activity assessed using a standardized measurement tool for RA- 4- No- Reason not specified
3250	Date when Disease activity assessed using a standardized measurement tool for RA

Aggregation of Encounters for a Given Patient

Denominator = pt-elig = max(enc-elig)
Numerator = pt-perf-met = max(enc-perf-met)
pt-perf-not-met = max(enc-perf-not-met) and not max(enc-perf-met)
Denominator Exclusion = pt-perf-excl = max(enc-den-excl) and not max(enc-perf-not-met) and not max(enc-perf-met)
Denominator Exception = pt-perf-exp = max(enc-den-exp) and not max(enc-perf-not-met) and not max(enc-perf-met)
pt-reported = max(enc-reported)

Aggregation of Patients for a Given Provider
eligible-instances = sum(pt-elig)
performance-met-instances = sum(pt-perf-met)
performance-not-met-instances = sum(pt-perf-not-met)
performance-exclusion-instances = sum(pt-perf-excl)
performance-exception-instances = sum(pt-perf-exp)
reported-instances = sum(pt-reported)
reporting-rate = reported-instances / eligible-instances
performance-rate = performance-met-instances / (performance-met-instances + performance-not-met instances

1.3. What are the dates of the data used in testing? 1/2013 to 12/2013

1/2017 to 12/2017

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

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

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

We recruited three testing sites that were geographically dispersed, included racial/ethnically diverse patient populations, and that used different electronic health record systems for collection of RA disease activity measures. We have summarized the geographic location and characteristics of the sites in Table 1 below.

Table 1. Geographic location, site characteristics and data sources used for rheumatoid arthritis disease activity quality measure.

Geographic Location	Site Characteristics	Data Source
Northeast United States	Large health system serving a largely *rural* population of over 2.6 million over 44 counties. The rheumatology clinics have over 24,000 patient visits per year. Within this system, rheumatology clinical encounters were analyzed.	*Rheum-PACER (Patient Centric Electronic Redesign).* This electronic, web-based platform pulls data from the health system's separate EMR as well as a patient touchscreen questionnaire completed at the start of each rheumatology visit, and provides both clinical staff and patients access to outcome measures at the point of care.
Western United States	Academic medical center located in an *urban* area that serves as a referral center in a geographic region of approximately 1 million residents. The rheumatology clinics have approximately 3000 patients visits per year.	*Epic-based electronic health record.* Documentation flowsheets were constructed within the Epic-based electronic record for collection of disease activity measures during routine rheumatology clinical care. Outcome measure data is available to both patients and clinicians in real-time within the electronic record.
Southeastern United States	Large community health system that serves both a *rural and urban* population in a statewide geographic region. The rheumatology clinics register over 20,000 visits annually.	*Cerner-based electronic health record*. Structured fields within the electronic record created to interface with an iPad-based patient data collection system. Use is being pilot-tested, preliminary data from automated electronic reports and also front-end electronic record reviews are provided.

For the signal-to-noise testing, we used data collected from outpatient rheumatology clinics that participate in the ACR's Rheumatology Informatics System for Effectiveness (RISE) registry. In the first quarter of 2017, 109 practices were fully connected to the RISE registry. The participating practices covered all regions of the country and represented a variety of practice settings: 27 solo practices, 78 group practices, two health systems, and two unknown settings. The practices used nearly 30 different EHR systems, including NextGen, eClinicalWorks, and Amazing Charts.

For testing purposes, the practices included in the signal-to-noise analysis were limited to those that were evaluated on measure performance from January 2017 through December 2017, which totaled 107. Of these 107, 27 (25%) practices were individual providers, while the other 80 (75%) were group practices or health

systems. Given the high percentage of individual providers also classified as individual practices, the analysis covers both individual- and practice-level results.

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

Data were analyzed at the individual patient level. *All* patients receiving care in rheumatology clinics in the Northeastern and Western health system were eligible for the denominator population if they met inclusion criteria, including ≥2 encounters for RA, being over age 18 years, and meeting these criteria over the measurement period of January 2013-December 2013. For the Southeastern site, only patients who were seen by the 2 providers participating in the site's pilot project were included.

For the front-end chart abstraction, a *simple random sample* was constructed for the Northeastern and Western sites. For the Southeastern site, the front-end chart abstraction included the entire denominator examined. The number of patients involved in the testing projects is included in Table 2 below.

Table 2. Patient characteristics of individuals with rheumatoid arthritis, by site, for quality measure testing studies.

Site	Total E-measure Patient Population (N)	Random Sample for Front-end EHR review (N)	Sex (% Female)
Northeastern site	1213	70	74%
Western site	400	119	83%
Southeastern site		34	

For the signal-to-noise testing, patients were included in the analysis if they were seen at one of the practices that met the practice inclusion criteria for Item 1.5 and if they met the patient inclusion criteria for the measure, including ≥2 encounters for RA, being over age 18 years, and meeting these criteria over the measurement period of January 2017 through December 2017. Across all sites, 94,872 patients met the inclusion criteria.

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.

For validity testing studies that involved a front-end electronic health record chart abstraction, a *simple random sample* of the eligible denominator population from the automated report generated by the e-measure was created for the Northeastern and Western Sites (see Table 2 for details). The characteristics of the random sample were similar to the denominator population.

For reliability testing, as noted above, we used physicians/practices reporting in 2017.

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

We do not routinely and uniformly collect social risk factors on all patients for this measure. Furthermore, we do not anticipate that measure reliability and validity would be impacted by social risk factors because the measure is a process measure, and therefore not risk-adjusted, and completion of the process at the core of this measure is important for all patients, regardless of patients' social status. Finally, the measure has been

tested and implemented with positive results without requiring social risk information, so we do not believe the analysis of social risk factors is required.

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*)

Please see section "2b2. VALIDITY TESTING" for testing results.

For signal-to-noise testing, data elements for this quality measure were extracted for the RISE registry from EHRs using computer programming, and therefore by virtue of automation, this process is repeatable (reliable); this was further verified during data element validation (described below). Data from the RISE registry included the number of patients and number passing the measure for each practice. With this, we can calculate pass rate and sample size for each practice, and we can compare variability in measure performance between practices. Because reliability depends on pass rate and sample size, it varies between practices.

Psychometricians use a rule of thumb of 90 percent for drawing conclusions about individuals. (*Hays RD, Revicki D. Reliability and validity (including responsiveness). In: Fayers P, Hays R, eds. Assessing Quality of Life In Clinical Trials. New York: Oxford University Press; 2005.; Adams, John L., The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corporation, 2009. https://www.rand.org/pubs/technical_reports/TR653.html.*) For binary measures, a tutorial by the RAND Corporation recommends fitting practices to a beta-binomial model. This can be done with the SAS Betabin macro (*Ian Wakeling - Qi Statistics. MACRO BETABIN Version 2.2 March 2005, www.qistatistics.co.uk*). This provides parameters a and b.

For the beta-binomial model, practice-to-practice variation = σ^2 = ab / ((a+b+1)*(a+b)^2).

Practice specific/measurement error for a binomial distribution = p*(1-p)/n; or when p = 1 or p = 0, substitute 3/n for p, by the rule of three.

Reliability = σ^2 / (σ^2 + p(1-p)/n), which represents the fraction of variance observed between practices not explained by practice specific variance.

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*)

Please see section "2b2. VALIDITY TESTING" for testing results.

For the signal-to-noise testing, each practice has a reliability score for the measure. The distribution of these practice-level scores is reported in Table 2a below.

Table 2a. Reliability scores for the Rheumatoid Arthritis: Assessment of Disease Activity measure among practices participating in the RISE registry, January 2017-December 2017.

Mean Reliability	Min Reliability	1st Quartile Reliability	Median Reliability	3rd Quartile Reliability	Max Reliability	Proportion of lowest quartile performers with reliability ≥0.9	Proportion of middle 50% performers with reliability ≥0.9	Proportion of highest quartile performers with reliability ≥0.9
0.97	0.48	0.97	1.00	1.00	1.00	NA†	0.93	1.00

†NA = not applicable; due to ties, there are no practices in this quartile.

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?*)

Data elements for this quality measure were extracted from EHRs using computer programming, and therefore by virtue of automation this process is repeatable (reliable); however, because data can be incorrect, testing focused on validity. Validity testing is outlined in detail below. Briefly, according to cutpoints that are commonly accepted (*Landis J, Koch G, The measurement of observer agreement for categorical data, Biometrics, 1977;33:159-174.)*, the overall Kappa in this study falls into the "near perfect" category. Validity testing results are discussed in more detail below.

Based on standard interpretations of reliability, these findings support strong reliability of the measure result. For the few extreme outliers with poor reliability, the poor performance is likely due to small case volumes and can, if needed, be addressed by flagging or suppressing any measure results based on very few observations.

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)*

Below, we discuss 2 different aspects of validity that are relevant to the proposed measure. These include: 1) Validity of critical data elements, obtained through comparison of automated e-measure data compared to a front-end total EHR data abstraction, as well as the validation performed during the RISE registry onboarding and yearly audit processes, and 2) Systematic assessment of face validity using the ACR's quality measure development process. *Reviewers are referred to materials elsewhere in the application that discuss the scientific literature supporting extensive validity studies of the measurement tools themselves, including their content and construct validity, responsiveness and comparability.*

1. Critical data element validity.

Data abstracted from randomly sampled patient records were used to calculate parallel forms reliability for the measure. Patient charts for abstraction were selected from visits for rheumatoid arthritis for adult patients with two or more face-to-face encounters for rheumatoid arthritis during the measurement period.

We examined whether EHR specifications and data exported electronically from the EHR were valid when compared to a front-end chart abstraction of the entire EHR by trained reviewers. From the population in which the e-measure was applied, we either reviewed all patient records (Southeastern site) or created a simple random sample (Northeastern and Western sites) for front-end abstraction. For the characteristics of sampled patients, please see Table 2 above.

Reviewers recorded relevant data elements using a structured data entry process. Overall performance rates using the automatically exported data as specified by the e-measure were compared to the front-end abstraction results by calculating a kappa coefficient, a statistical measure of inter-rater agreement.

To ensure data integrity, additional measures were taken. For example, one site was instructed to blind reviewers to the results of the automated report. Each record underwent front-end review by two separate reviewers, and conflicts in this front-end data were adjudicated by the project lead investigator (conflicts N=2 out of 119, front-end inter-rater reliability 0.97, range 0.92 to 1.00).

For the QDM data element "Diagnosis: Rheumatoid Arthritis" front-end chart review found disagreement in 1.8% of cases compared to the automated report. These instances resulted from the provider improperly coding the patient's diagnosis as RA, when in fact the patient had another diagnosis, often with an inflammatory arthritis component (e.g. mixed connective tissue disease). These data are consistent with the scientific literature in which the validity of case definitions for RA using related automated algorithms have been examined (*Chung CP, A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. Vaccine. 2013 Dec 30;31 Suppl 10:K41-61; Carroll RJ et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012 Jun;19(e1):e162-9*; *Liao KP, Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res. 2010 Aug;62(8):1120-7*). Conclusions from this literature are that algorithms, such as the one used here, in which more than one code for RA is required, including a diagnosis from a rheumatologist, have good sensitivity and specificity.

For the QDM data element "Risk Category Assessment: Rheumatoid Arthritis Disease Activity Measurement Tools (result)" disagreement was found in 2.2% of the testing sample compared to the automated report. In these instances, the patient did not meet the threshold for "Risk Category Assessment: Rheumatoid Arthritis Disease Activity Measurement Tools (result)" during at least 50% of Encounters Performed for RA during the measurement year. This was the result of specific problems with structured data that were later addressed. For example, at one site, providers could enter disease activity measure scores outside of an encounter, which led to a mismatch between the automated and front-end review results. Scores are now linked to encounters, so this problem was resolved.

As noted in section 1.2, this measure has been implemented in the ACR's RISE registry. RISE uses computer programming to extract data from the EHR systems of participating providers, analyze the data and provide feedback through an analytic dashboard on a provider's performance on this measure. Through the implementation process, providers must confirm that all data used to calculate the measure performance is accurate and valid. The dashboard is updated on a monthly basis and allows providers to track their performance over time. This allows providers to regularly assess the accuracy of their measure performance score. If providers discover any inconsistencies, they work directly with RISE registry technical experts to identify and correct the source of the issue.

While ACR is transparent about the specifications, this is functionally a registry measure, similar to STS' NQF-endorsed measures that cannot be reproduced by other entities, and thus the quality of the output (and the validity of normalized values) is performed through iterative work between the practices, the registry tech vendor and our third-party data analytic centers that review the data collected by the vendor during set-up of the practices and on a regular basis.

Furthermore, the RISE dashboard allows providers to see how their performance on each quality measure, including the Rheumatoid Arthritis: Assessment of Disease Activity measure, compares to the average performance of all RISE providers. During the onboarding process, practices not only evaluate their own data to ensure that each element is accurate and valid; they also evaluate their performance against the registry average. Because all practices in RISE go through the same onboarding process, practices are able to verify that any difference in their measure performance as compared to the registry average is due to differences in quality of care.

The RISE registry also conducts yearly audits to verify the accuracy of the patient data extracted from the EHR systems of a random sample of participating practices. The most recent audit was conducted in 2018 on data from January 2017 to December 2017. Random sampling technique was used for a sample size of 13 TIN/NPI combinations. For each TIN/NPI sample, a minimum of 40-50 patients were reviewed for audit purposes. Providers reviewed and reported back on the accuracy of data for all reportable measures applicable to the patient, including data relevant to this measure.

2. Systematic assessment of face validity. Systematic assessment of face validity was performed using a multi-stakeholder expert panel that formally rated validity of the proposed measure using a scale based on the RAND Appropriateness Method. *Panelists participated in an open and transparent process in which they were specifically asked to address whether the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*.

The American College of Rheumatology has worked for the last several years to develop a rigorous measure development process that leverages the considerable investment in producing guidelines and also input from stakeholders throughout the health care system in the area of rheumatoid arthritis (RA). *The following information is provided to place the Expert Panel ratings, used to assess face validity, in context*. The major elements of the measure development process are listed here. Reviewers are referred to materials in the supplemental appendix for further details.

- First, the ACR assembled a **Working Group** of 7 experts in RA, quality measurement, and health services research meeting its conflict of interest policies (requiring that a majority of group members, including the principal investigator, have no links to any company or commercial entity that makes a drug, device or product in the area of RA). The Work Group was tasked with drafting potential quality measures based on 2012 ACR Guidelines for the management of RA (*Singh JA, Furst DE, Bharat A et al. 2012 update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken). 2012 May;64(5):625-39*). Measures were drafted in an iterative fashion over a period of six months.
- Preliminary measures were presented to a separate multi-stakeholder **Expert Panel** of 16 for formal ratings. The group was comprised of patients with RA, practicing rheumatologists whose primary responsibility is patient care, an orthopedic surgeon nominated by the American Academy of Orthopedic Surgery, an Internal Medicine specialist nominated by the American College of Physicians, a member of the American Rheumatology Health Professional's Association, a payer representative (a Medical Director for a large public payer program), and methodological experts with expertise in quality measure development. For each measure, the panel was asked to review the scientific evidence and vote prior to meeting. These results were then presented to the panel and a facilitated discussion using initial ratings was undertaken during a meeting. Members voted again after deliberating. Results were analyzed according the RAND Appropriateness Method (mean scores of 7-9 indicate good agreement if criteria for disagreement are absent; *see Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, editors. Methodology perspectives. Rockville (MD): US Department of Health and Human Services; 1994. p. 59–70*). Panel ratings on the measure are provided below. Table 3 summarizes the results of the rating procedure. ***The median score for validity was 9 (indicating excellent validity).***

Table 3. Data from the American College of Rheumatology's Rheumatoid Arthritis Quality Measures Project Expert Panel Rating Process for Disease Activity Measure.[1,2]

Median score for validity	Median score for feasibility	# of raters with validity score ≤ 3	# of raters with validity score ≥ 7	# of raters total	% invalid (score ≤ 3)
9	7	1	11	14	7.14%

1. *Panelists were provided with the following instructions*: "Your validity ratings should reflect whether you believe that the measure can be used to reflect the quality of care for RA. Questions to consider in determining your validity ratings should include:

a. Is there adequate scientific evidence or professional consensus to support the indicator?

b. Are there identifiable health benefits to patients who receive care specified by the indicator?

c. Based on your professional experience, would you consider providers with significantly higher rates of adherence to the indicator higher quality providers?

d. Are the majority of factors that determine adherence to the indicator under the control of the physician or health care system?"

2. *Measure scale definitions*: For validity, 1=definitely NOT valid to 9=definitely valid; for feasibility, 1=definitely NOT feasible; 9=definitely feasible.

- In addition to the formal validity assessment by experts, additional vetting was performed in several ways. First, the ACR requested **public comment** on the measure, publicizing the comment period through email communication with ACR members and communicating with the leadership of other stakeholder groups. Public comments were reviewed and did not identify any additional issues concerns with the measure.
- Finally, the **ACR Quality Measures Subcommittee, ACR Quality of Care Committee** and **ACR Board of Directors** approved the measures.

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

1. Critical data element validity.

Kappa *Overall*, Range, % Agreement: **0.81** (0.73 to 0.89), 0.91

Kappa, Range, % Agreement Denominator: 0 (0, .97) 98.2%

Kappa, Range, % Agreement Numerator: **0.84**, 0.77 to 0.91, 92.2%

Kappa, Range, % Agreement Exceptions: **1.00** (1.0 to 1.0), 100%*

*100% agreement that there are no exceptions

Recommended guidelines for interpreting Kappa values from the National Quality Forum's Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

Kappa values range between 0 and 1. 0 and are interpreted as degree of agreement beyond chance. By convention, a kappa > .70 is considered acceptable inter-rater reliability, but this depends on the researcher's purposes[23]	
0	No better than chance
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.0	Almost perfect[23]

Because instances of agreement dominated, the denominator Kappa was zero. The instance of 0 for the denominator is an example of the limitation of the Kappa statistic. A kappa of zero can be obtained even though agreement is very high due to one classification category dominating.

(See http://www.ajronline.org/doi/abs/10.2214/ajr.184.5.01841391 for full details).

Please see above section for details of additional validity testing results.

Table 3a below contains the results from the registry audit conducted in 2018.

Table 3a. Results of RISE registry audit of data from January 2017-December 2017.

Number of NPI/TIN audited	Number of Patients	Expected count of Responses	Number of Correct Responses	Number of Incorrect Responses	% Success	% Fail
13	644	698	684	14	97.99%	2.01%

2. Systematic assessment of face validity.

Table 3. Data from the American College of Rheumatology's Rheumatoid Arthritis Quality Measures Project Expert Panel Rating Process for Disease Activity Measure.[1,2]

Median score for validity	Median score for feasibility	# of raters with validity score ≤ 3	# of raters with validity score ≥ 7	# of raters total	% invalid (score ≤ 3)
9	7	1	11	14	7.14%

1. *Panelists were provided with the following instructions*: "Your validity ratings should reflect whether you believe that the measure can be used to reflect the quality of care for RA. Questions to consider in determining your validity ratings should include:

a. Is there adequate scientific evidence or professional consensus to support the indicator?

b. Are there identifiable health benefits to patients who receive care specified by the indicator?

c. Based on your professional experience, would you consider providers with significantly higher rates of adherence to the indicator higher quality providers?

d. Are the majority of factors that determine adherence to the indicator under the control of the physician or health care system?"

2. *Measure scale definitions*: For validity, 1=definitely NOT valid to 9=definitely valid; for feasibility, 1=definitely NOT feasible; 9=definitely feasible.

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?*)

Critical data element validity. The kappa statistic of 0.81 for overall performance indicates high agreement between the automated report and the front-end chart abstraction. Individual data elements were found to be highly reliable.

Manual audit validity testing results in a random sampling of practices indicated a very high (98%) accuracy.

Systematic assessment of validity. Ratings by a multi-stakeholder group in which the RAND/UCLA rating scale was applied found excellent validity of this measure, with a mean score of 9, and no disagreement.

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 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

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

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

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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

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

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

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

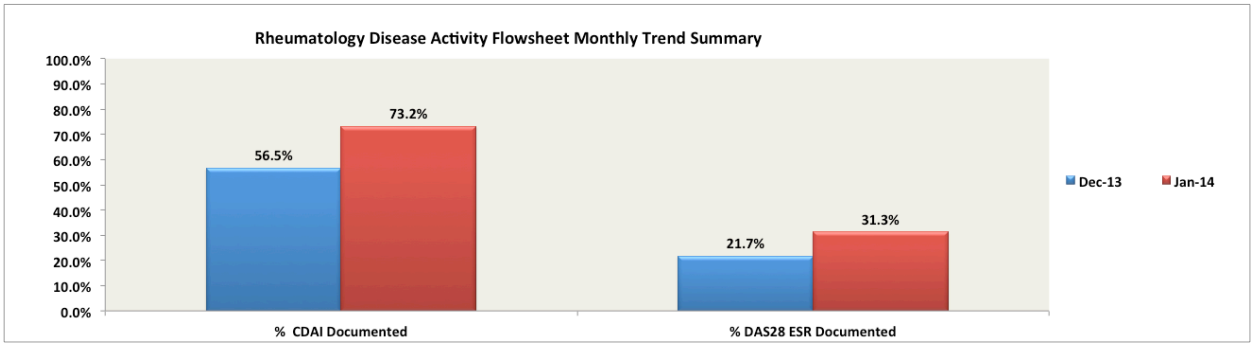
2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat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

Performance varied both between sites and between providers at sites. Differences between providers in use of validated disease activity assessments in RA reflects a meaningful gap in quality, based on qualitative feedback from clinicians and statistical analysis of the data. For example, at the Western site, performance at the individual provider level for disease activity measurement varied significantly (range 0 to 100%, mean 65%, SD 35%). This is consistent with variation in disease activity measurement reported in the scientific literature (*Adhikesavan LG et al. American College of Rheumatology quality indicators for rheumatoid arthritis: benchmarking, variability, and opportunities to improve quality of care using the electronic health record. Arthritis Rheum. 2008 Dec 15;59(12):1705-12.)* and from data available from the ACR's Rheumatology Clinical Registry on an earlier version of this quality measure (*Yazdany J et al. Uptake of the American College of Rheumatology's Rheumatology Clinical Registry (RCR): Quality Measure Summary Data". Annual Scientific Meeting. American College of Rheumatology. Reed Convention Center, Washington, DC. 27 October 2013. Arthritis Rheum abstract supplement*).

Statistical testing using regression models, with weights applied to account for the fact that providers have different numbers of eligible patients confirmed significant variation ($p<0.001$). Performance ranged from 35-61% in the sites that have established a workflow to collect disease activity measures.

It should be noted that although performance was found to vary between providers, performance on this measure appears to have potential to improve. As an example, monthly performance reports fed back to providers as part of a quality improvement project at one of the testing sites resulted in significant improvement between December 2013 and January 2014, see Figure 1. In December, 6 of 15 providers did not meet the performance threshold for performing disease activity assessments (50% or more of RA encounters), by January of 2014, this number had decreased to only a single provider.

Figure 1. Example of Performance Improvement on Rheumatoid Disease Activity Measure at Testing Site over One Quality Improvement Cycle.



In addition, some differences in performance across sites (i.e. between the Southeastern testing sites and other sites) is attributed to whether or not clinicians are entering structured data as part of the current workflow. Sites that have established workflows to capture these data in structured fields had higher

performance on the e-measure compared to the site where this workflow is still in the early stages. Implementation of this e-measure in the United States will require changes in clinical workflow for many practices, which may require customized solutions at individual sites. Support by EHR vendors in providing these tools will likely speed implementation.

We also evaluated the variation in measure performance in 2017 among 107 RISE practices, representing 98.2% of all practices fully enrolled in RISE at the beginning of 2017.

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*)

Data from the sample are provided above.

National benchmarking data for this e-measure are currently not available. However, an earlier version of this measure that was part of the Physical Quality Reporting System (PQRS) since 2008, requiring providers to measure disease activity at least once per year and categorize it as remission, low, moderate or high, found suboptimal performance. Data reported through the ACR's Rheumatology Clinical Registry (RCR) indicate that performance was 43.4% in 2011, improving to 54.4% in 2012 (*Yazdany J et al. Uptake of the American College of Rheumatology's Rheumatology Clinical Registry (RCR): Quality Measure Summary Data". Annual Scientific Meeting. American College of Rheumatology. Reed Convention Center, Washington, DC. 27 October 2013. Arthritis Rheum abstract supplement*).

Table 4. Variation in performance on Rheumatoid Arthritis: Assessment of Disease Activity measure in the RISE registry, January 2017-December 2017.

Practices	Total Denominator	Mean Denominator	Denominator range	Total Numerator	Mean Numerator	Numerator Range	Average Practice Performance (%)	25th, 50th, 75th, 100th percentile
107	94872	886.65	18-4017	50080	468.04	0-3932	43.91%	0.40, 42.96, 80.49, 100

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?*)

Experience with an earlier version of this measure in the PQRS program, which found that the national mean among participating providers for performing even a yearly disease activity assessment was only 54% in 2012 (see above), as well as testing of the current e-measure suggest that there are meaningful differences in performance across providers. Importantly, data from an ACR benchmarking survey suggests that 69.6 percent of U.S. rheumatologists currently perform any form of disease activity assessment in clinical practice (ACR Benchmarking Survey, 2013).

The results demonstrate both wide variation and a continued need for improvement in performance overall given that the average performance in 2017 was 43.91%; the drop in average success from prior assessments likely reflects both changing demographics and a shift from non-EHR-based measure versions used in the past. Optimal clinical performance for this measure should be 100%, as this measure reflects the ACR guidelines for care of RA patients and what is required of providers to adequately assess the progress of their patients'

disease in an empirical manner. An average measure score under 44% (and a 75th percentile of 80%) supports an ongoing opportunity for improvement in performance.

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and 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*)

Identification of missing data was included as part of the earlier critical data element validity testing described in section 2b1.

With the RISE registry, there is no missing data. As described in section 1.2, during the implementation process, providers work with the registry's technical experts to review the data elements necessary for measure performance calculations and direct the technical team on how to find those data elements in the practice's EHR system. The technical team is them able to extract the necessary data from both structured and unstructured fields. This ensures that accurate measure performance can be calculated no matter how the information is documented (in free text or as a scanned pdf).

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*)

Not applicable – there were no missing data in our earlier testing. As noted above, the RISE registry's data abstraction approach ensures there is no missing data. See 2b6.3.

However, because it may be of interest to reviewers, we provide additional information beyond that required for measure testing to understand the potential for data to be missing on this PRO.

Because all of the proposed disease activity measures (CDAI, SDAI, DAS, RAPID, PAS) require a patient-reported component, patient non-response may lead to missing data and inability to capture a disease activity

score. There are no procedures for handling missing data because we found missing-ness to be a rare occurrence during our systematic assessment, described below.

Systematic examination of missing data. Clinicians are able to select from a list of validated disease activity measures. Taking the example of the CDAI, SDAI and DAS, measures that require a "patient global assessment of disease activity" that is part of a composite score, we examined missing data from the patient-reported component of the score. "Patient globals" are collected on a visual analog scale, available in multiple languages. Because these assessments simply require the patient to place an "x" on a line, this measure is appropriate for use in very low literacy populations. Nevertheless, some missing data on this component occurs in routine clinical practice. *One of our testing sites examined this issue in more detail.* The site serves a multi-ethnic population that is socioeconomically diverse and has variable health literacy. Medical assistants administer a patient global assessment questionnaire upon patient registration in the clinic in the patient's primary language, including in English, Spanish or Chinese. Among over 400 individuals with RA, 2 (<1%) of individuals declined completing the forms during clinical encounters. *Missing data was therefore found to be a rare occurrence*.

The findings of our testing study are consistent with validation studies in the literature, which include systematic assessments of respondent burden and missing data. A summary of this literature can be found in the following paper and its appendices: *Anderson J et al. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken). 2012 May;64(5):640-7*.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

As described above, because missing data were a rare occurrence, no additional procedures or sensitivity analyses were undertaken to evaluate missing data; missing data are not expected to influence over performance, particularly given that the performance threshold is ≥ 50%. *Providers caring for low-literacy or at-risk populations have the option of selecting a disease activity measure that is appropriate to their setting and specific patient population.*

Furthermore, because of the method of data mining used to calculate measure performance in the RISE registry, the absence of a necessary data element, such as a lab test, a medication or a disease activity assessment, is not indicative of missing data. Rather, it indicates that the provider did not perform the expected action.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in a combination of electronic sources

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

As noted in S.3.2., the ACR made a conscious decision to move away from an eCQM in order to provide the most flexible route to electronic health record data-based measurement and avoid forcing individual practitioners to change their workflow and documentation to satisfy requirements for HQMF specifications. The ACR will continue to monitor developments in coding and HQMF specifications to determine if the updates would provide the necessary flexibility to make this measure an eCQM.

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

Attachment: RA_Feasibility_Survey_Responses_-_Data_Element_Scores-635291967610444146.xls

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.

Measurement of RA disease activity using a standardized, validated instrument requires significant changes to current clinical workflow for many practices. A range of options were provided in the measure since 1) there is good scientific evidence suggesting comparable validity of several measures, which are endorsed by the ACR as recommended measures, and 2) feasibility of using different measures depends greatly on the U.S. practice setting. Rheumatologists practice in a great variety of settings, including solo clinical offices, single and multi-specialty group practices, and academic and large group settings. Resources available for disease activity measures vary between practices, and the consensus process for this measure took this into account. For example, academic medical centers may have same-day laboratory information available, allowing calculation of a Disease Activity Score 28 (DAS 28) or Simplified Disease Activity Index (SDAI). Small or rural practices often do not have access to same day laboratory results and may not have adequate support staff to implement the more complex workflow required for composite measures such as the Clinical Disease Activity Assessment (CDAI). In these settings, the Routine Assessment of Patient Index Data (RAPID3) or Patient Activity Scores (PAS) are both valid and feasible to implement. There is a large body of research spanning many decades regarding these outcome measures in RA. In a large, national effort that involved many stakeholders, the ACR has summarized information on use of these measures in clinical practice. This manuscript includes

information on time to collect each measure and feasibility based on practice setting. It is important to note that the recommended measures all have cut-offs for remission and low, moderate and high disease activity.

Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, Saag KG, O´Dell JR, Kazi S. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken). 2012 May;64(5):640-7.

Yazici Y, Bergman M, Pincus T. Time to score quantitative rheumatoid arthritis measures: 28-Joint Count, Disease Activity Score, Health Assessment Questionnaire (HAQ), Multidimensional HAQ (MDHAQ), and Routine Assessment of Patient Index Data (RAPID) scores. J Rheumatol. 2008 Apr;35(4):603-9.

Anderson JK, Zimmerman L, Caplan L, Michaud K. Measures of rheumatoid arthritis disease activity: Patient (PtGA) and Provider (PrGA) Global Assessment of Disease Activity, Disease Activity Score (DAS) and Disease Activity Score with 28-Joint Counts (DAS28), Simplified Disease Activity Index (SDAI), Clinical Disease Activity Index (CDAI), Patient Activity Score (PAS) and Patient Activity Score-II (PASII), Routine Assessment of Patient Index Data (RAPID), Rheumatoid Arthritis Disease Activity Index (RADAI) and Rheumatoid Arthritis Disease Activity Index-5 (RADAI-5), Chronic Arthritis Systemic Index (CASI), Patient-Based Disease Activity Score With ESR (PDAS1) and Patient-Based Disease Activity Score without ESR (PDAS2), and Mean Overall Index for Rheumatoid Arthritis (MOI-RA). Arthritis Care Res (Hoboken). 2011 Nov; 63 Suppl 11:S14-36.

Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, Saag KG, O´Dell JR, Kazi S. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken). 2012 May;64(5):640-7.

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

N/A

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program	Payment Program MIPS https://qpp.cms.gov/mips/overview Quality Improvement (external benchmarking to organizations) The RISE Registry www.riseregistry.org Quality Improvement (Internal to the specific organization) The RISE Registry www.riseregistry.org

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

Program: Merit-based Incentive Payment System
Sponsor: Centers for Medicare and Medicaid Services
Purpose: MIPS was designed to tie payments to quality and cost-efficient care, drive improvement in care processes and health outcomes, increase the use of healthcare information, and reduce the cost of care.
Geographic area: United States
Number and percentage of entities and patients: Per the most recent numbers provided by CMS*, approximately 3,550 rheumatologists across the country (and 100% of their patients) are eligible for MIPS reporting
Level of measurement: provider or practice, depending on whether they report as an individual or group
Setting: Non-hospital-based rheumatology practices enrolled in Medicare the exceed the low-volume threshold
* Page 374: https://www.govinfo.gov/content/pkg/FR-2017-11-16/pdf/2017-24067.pdf
Program: The Rheumatology Informatics System for Effectiveness (RISE) registry
Sponsor: American College of Rheumatology
Purpose: To help prepare rheumatologists for the significant challenges of a rapidly changing healthcare environment, including adapting to new payment and delivery models, meeting evolving certification requirements, and using EHR data to assess quality of care.
Geographic area: United States
Number of entities and patients: As of January 3, 2019, 937 rheumatology providers participated in RISE, representing 1,787,394 patients
Level of measurement: provider and practice
Setting: Solo practice, single-specialty group practice, multi-specialty group practice

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

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.*)

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.

For information on feedback from those being measured during measure development, please refer to the validity testing in section 2.

For implementation, those being measure are deeply involved in the process. Measure performance is shared with rheumatology providers via the ACR's RISE registry. Participating providers work closely with the registry technology vendor to ensure data is being extracted from their EHR correctly and portrayed accurately via the registry's analytic dashboard. Through the RISE dashboard, providers are able to see their individual overall performance on the measure, their practice's overall performance on the measure, and the average performance of all RISE users on the measure. Each provider is also able to drill down into their measure performance to see the patients who qualify for the denominator and the numerator. Furthermore, providers have direct access to the human readable measure specifications in the dashboard. If they have any questions or concerns about how the measure is being calculated or the specifications in general, they are able to

contact both ACR staff and the registry technology vendor staff directly. This allows providers the ability to confirm the accuracy of their measure performance, review how their own practices impact their measure performance, and get any questions on measure interpretation answered directly by the measure owner.

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.

The analytic dashboard for all RISE providers is updated every month following the most recent data extraction. All providers have constant access to their analytic dashboard to review the measure specifications and their measure performance. ACR and vendor staff are available during regular business hours to answer their questions over the phone or via e-mail.

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.

RISE users communicate directly with the registry technology vendor and ACR staff over the phone and via e-mail.

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

When communicating with staff, they have said that this and the other measures included in the registry to be very helpful in understanding the quality of care they provide patients. When a provider first joins the RISE registry, most often they note that they expected higher performance on their measures. However, through their work with the registry technology vendor and the analytic dashboard, they are able to see an objective analysis of their data and realize that they are not providing as high of quality care as they assumed. The other most common feedback received on this measure is focused on ways to identify the various data elements in the measure. For example, a provider may use a different tool than approved for use in the measure or document a lab result in a different way than expected.

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

As far as we are aware, this measure has only been implemented in the RISE registry until recently. This measure was previously used by RISE participants for PQRS reporting. However, when CMS transitioned to MIPS, they denied inclusion of this measure as a QCDR measure because they said it was too similar to a less stringent QPP measure. We have since updated the QPP measure for the 2019 reporting year to conform to the more stringent requirements of this measure. Because of this, we have not received feedback from other entities. However, we will have the opportunity to begin doing so in 2020.

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

As noted, RISE providers have direct communication with the registry vendor and ACR staff. They are able to ask questions and share concerns directly with the ACR and receive prompt feedback. As needed, ACR staff are able to take questions and concerns to a team of rheumatology volunteers with expertise in quality measurement. Feedback from ACR and the quality measure experts is then used to improve the guidance on quality measure implementation for both the registry technology vendor and the provider.

Improvement

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

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

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

The decrease in performance between 2016 Q2 and the end of 2017 reflects persistent low performance of routinely collecting and documenting disease activity among US rheumatologists. The prior increasing performance likely reflected an early adopter phenomenon, where early RISE adopters were more likely to have systems in place to collect a range of data elements, including disease activity, and they were receiving quarterly results allowing them to implement improvements. The over doubling of the number of practices in RISE between the two time periods (44 to 107), many in response to the MACRA legislation, probably reflects a more generalizable group of US rheumatology practices. Furthermore, during this time, CMS shifted from providers reporting on this measure to a less stringent version which only required assessment of disease activity once a year. The current measure, as specified, is only now being implemented in MIPS for the 2019 reporting year. The variation in results indicates continued need for assessing performance on this measure, especially as more practices continue to join RISE and providers are once again held to this higher standard as supported by RA treatment guidelines.

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.

As noted in S.3.2, we found that many providers were documenting key aspects of the measure data elements in free text or other non-standardized formats. Only a portion of providers have laboratory data and/or prescription data integrated into their outpatient electronic health record, further complicating the ability to pull HQMF-formatted specifications.

We are unaware of any negative or unintended impacts on patients due to measurement.

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

We received positive feedback from several participating providers. This included both the benefits of better understanding provider variation within practices as well as identification of higher-risk patients such as those with frequent disease flares.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

No

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

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

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

5b. Competing Measures

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

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

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-635294351903102622.xlsx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American College of Rheumatology

Co.2 Point of Contact: Rachel, Myslinski, rmyslinski@rheumatology.org, 404-633-3777-824

Co.3 Measure Developer if different from Measure Steward: American College of Rheumatology

Co.4 Point of Contact: Rachel, Myslinski, rmyslinski@rheumatology.org, 404-633-3777-824

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.

Jinoos Yazdany, MD, MPH

University of California San Francisco

Mark Robbins, MD

Harvard Vanguard Medical Associates

Sonali Parekh Desai, MD

Diane V. Lacaille, MD, FRCPC, MHSc

Arthritis Research Center Canada

Gabby Schmajuk, MD

University of California San Francisco

Eric Newman, MD

Geisinger Medical Center

Jasvinder Singh, MD

University of Alabama Birmingham

Tuhina Neogi, MD

Boston University

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

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

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

Ad.6 Copyright statement: Copyright (c) 2013, American College of Rheumatology

Ad.7 Disclaimers: All materials are subject to copyrights owned by the College. The College hereby provides limited permission for the user to reproduce, retransmit or reprint for such user´s own personal use (and for such personal use only) part or all of any document as long as the copyright notice and permission notice contained in such document or portion thereof is included in such reproduction, retransmission or reprinting. All other reproduction, retransmission, or reprinting of all or part of any document is expressly prohibited, unless the College has expressly granted its prior written consent to so reproduce, retransmit, or reprint the material. All other rights reserved.

CPT(R) contained in the Measure specifications is copyright 2004-2013 American Medical Association.

LOINC(R) copyright 2004-2012 Regenstrief Institute, Inc.

This material contains SNOMED Clinical Terms(R) (SNOMED CT[R]) copyright 2004-2012 International Health Terminology Standards Development Organisation.

ICD-10 copyright 2012 World Health Organization. All Rights Reserved.

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

Ad.8 Additional Information/Comments: