**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Measure Title**: Rheumatoid Arthritis: Assessment of Functional Status

**Date of Submission**: Click here to enter a date

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process **(Structured assessment of outcome)** |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

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

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 functional status 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. |

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

Data were analyzed at the individual patient level. *All* patients receiving care in rheumatology clinics in the Northeastern and Western health systems 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 are 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 Population  (N) | Random Sample for Front-end EHR review  (N) | Sex  (% female) |
| Northeastern site | 1206 | 70 | 74% |
| Western site | 400 | 119 | 83% |
| Southeastern site |  | 34 |  |

**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**.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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.

**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.

As noted above, because this is an e-measure, data is by definition reliable. The focus of testing was on data element and overall score validity. See more detailed information under validity below.

**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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. 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*)

**2b2.2. 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 3 different aspects of validity that are relevant to the proposed measure. These include: 1) Validity of the performance measure score, obtained through comparison of automated e-measure data compared to a front-end total EHR data abstraction and 2) Validity of critical data elements, and 3) 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. Performance measure score 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 abstraction were adjudicated by the project lead investigator (conflicts N=0 out of 119, front-end inter-rater reliability kappa=1.0).

**2. Critical data element validity.**

**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 “Intervention Performed: Rheumatoid Arthritis Functional Status Assessment”** disagreement was found between the automated report and the front-end abstraction in 7.6% of the testing sample. These instances were due to data not being captured by the automated EHR report because they did not reside in structured fields.

**3. 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 to 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 8 (indicating excellent validity).***

**Table 3. Data from the American College of Rheumatology’s Rheumatoid Arthritis Quality Measures Project Expert Panel Rating Process for Functional Status 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) |
| 8 | 8 | 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.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

**1. Performance measure score validity results**

Sample Size: 223

Kappa Overall, Range, % Agreement: 0.95, 0.91 to 0.99, 97.3%

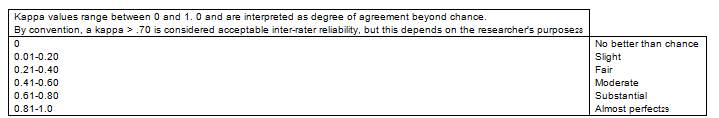
Kappa, Range, % Agreement Denominator: 0, 0 to 0.97, 98.2%

Kappa, Range, % Agreement Numerator: 0.96, 0.92 to 1.0) 98.2%

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



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).

**2. Critical data element validity.** Please see above section for details of validity testing results.

**3. 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 Functional Status 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) |
| 8 | 8 | 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.

**2b2.4. 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?*)

**E-measure validity testing**. The kappa statistic of 0.95 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.

**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 8, and no disagreement.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. 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*)

**2b3.2. 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*)

**2b3.3. 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*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

**2b4.2. If an outcome or resource use 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**.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)

**2b4.4. What were the statistical results of the analyses used to select risk factors?**

**2b4.5. 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*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b4.9. Results of Risk Stratification Analysis**:

**2b4.10. 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*)

**2b4.11.** **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*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. 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 in collection of patient-reported outcomes (PROs) such as functional status reflect differences in performance, 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 functional status measurement varied significantly (range 0 to 100%, mean 63%). This is consistent with variation in functional status 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 accounting for repeated observations and with the provider as a random effect (to account for the fact that providers have different numbers of eligible patients) confirmed that there was significant inter-provider variation (p<0.001). Performance ranged from 62-92% in the sites that have established a workflow to collect disease activity measures.

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.

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, 3 of 15 providers did not meet the performance threshold for performing functional status assessments, by January of 2013, this number had decreased to only a single provider.

**2b5.2. 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 functional status (using any means, rather than a validated assessment as is being recommended here) at least once per year are available. Data reported through the ACR’s Rheumatology Clinical Registry (RCR) indicate that performance was 69.6% in 2011, improving to 86.6% 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*).

**2b5.3. 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, 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 61.9 percent of U.S. rheumatologists currently perform any form of functional status assessment in clinical practice (ACR Benchmarking Survey, 2013).

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of 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*)

**2b6.2. 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*)

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of 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*)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. 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*)

Not applicable – there were no missing data in our testing.

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 functional status measures (HAQII, MDHAQ, PROMIS PF) are patient-reported outcomes (PROs), patient non-response may lead to missing data and inability to record a score, leading to low performance. There are no procedures for handling missing data as missing-ness was relatively uncommon in our testing and the requirement for functional status assessment once per year had face validity in our Expert Panel process.

**Systematic examination of missing data**. Clinicians are able to select from a list of functional status measures (HAQII, MDHAQ, PROMIS PF), including measures that have been validated for low-literacy populations (Grade 6 reading level) and that are available in multiple languages. To understand how often such data are missing in clinical encounters, we examined data from one of our testing sites in more detail. This site has a diverse patient population, including at-risk populations with low health literacy. At this site, the PROMIS PF10 functional status status is administered to patients in their native language (English, Spanish or Chinese). Out of 915 encounters over the measurement year, we have 58 instances **(6.34%)** of visits where a PROMIS score was not recorded but a “patient global assessment (PGA)”was recorded. Because the PGA and PROMIS scores are administered simultaneously in the site’s workflow, and PGAs are completed by nearly all patients because they do not require the patient to read or complete a survey, discrepancies between the 2 scores represent a PROMIS PF “non-responder” rate. Non-response to this PRO was therefore approximately 6.3%.

The findings of our testing study are consistent with validation studies in the literature, which include systematic assessments of respondent burden. A summary of this literature can be found in the following sources: *Maska L et al. Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment Questionnaire II (HAQ-II), Improved Health Assessment Questionnaire (Improved HAQ), and Rheumatoid Arthritis Quality of Life (RAQoL). Arthritis Care Res (Hoboken). 2011 Nov;63 Suppl 11:S4-13 and www.nihpromis.org‎*.

**2b7.2. 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*)

As described above, because missing data occurred in <10% of cases, no additional procedures or sensitivity analyses were undertaken to deal with missing data. Missing data are not expected to significantly influence overall performance in most settings, particularly given that the performance threshold is performing functional status assessment once per year.

Providers caring for a very large number of individuals with low health literacy may have lower performance. This will require further testing to unravel. However, a strength of the current measure is that it allows providers to select a functional status measure that is best suited to their setting and specific patient population. In addition, new technologies, including administering PROs using touchscreens that are voice-enabled may resolve some of these issues over time.

**2b7.3. 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*)

See above.