**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 2522e

**Measure Title**: Rheumatoid Arthritis: Tuberculosis Screening

**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 |  |

|  |
| --- |
| **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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **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 (including clinical and social risk factors) that influence the measured outcome 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.   **2b4.** 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.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **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.** 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.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: Click here to describe | other: Click here to describe |

**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**

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

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: Tuberculosis Screening measure can be seen in Figures 2a and 2b.

**Figure 2a. Flowchart of calculation for the Rheumatoid Arthritis: Tuberculosis Screening measure**

Figure 2a. Flowchart of calculation for the Rheumatoid Arthritis: Tuberculosis Screening measure
The measure flowchart show the process of the programming to determine inclusion in and calculation of measure performance for each measure. It does the following:
- Reads the encounter
- Checks to ensure patient is 18 years or older
- Checks the encounter is in the reporting period
- Checks for a diagnosis of RA during measurement period
- Checks for newly prescribed biologic DMARD during the measurement period.
- Checks for documentation of screening for TB in preceding 12 months
- Determines if performance was met or not

**Figure 2b. Supplement to flowchart of calculation for the Rheumatoid Arthritis: Tuberculosis Screening measure**

Figure 2b. Supplement to flowchart of calculation for the Rheumatoid Arthritis: Tuberculosis Screening measure
The supplement to the measure flowchart shows the element name and the associated ID numbers to identify the element in the RISE data dictionary. It also shows the formulas used to calculate the various elements in the calculation of measure performance.

**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 all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: 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 one large testing site to examine the TB testing measure. We have summarized the geographic location and characteristics of the site in Table 1 below.

**Table 1. Geographic location, site characteristics and data source used for Tuberculosis measure testing.**

|  |  |  |
| --- | --- | --- |
| 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. |

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: 28 solo practices, 77 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 105. Of these 105, 26 (25%) practices were individual providers, while the other 79 (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 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 at this site 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 front-end chart abstraction, a *simple random sample* was constructed for analysis. The number of patients involved in the testing project 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 | 87 | 66 | 69% |

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, 9,943 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 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 = σ2 = 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 = σ2 / ( σ2 + 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.

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.

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: Tuberculosis Screening 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** |
| 0.77 | 0.12 | 0.48 | 0.95 | 0.98 | 1.00 | 0.50 | 0.73 | 0.56 |

**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 is excellent. 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.** Validity of critical data elements. 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. 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.

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: Tuberculosis Screening 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 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 9 (indicating excellent validity).***

**Table 3. Data from the American College of Rheumatology’s Rheumatoid Arthritis Quality Measures Project Expert Panel Rating Process for Tuberculosis Screening 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 | 8.5 | 0 | 14 | 14 | 0% |

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.**

Sample Size: 66

Kappa Overall, Range, % Agreement: 1.00, 1.0 to 1.0, 100%

Kappa, Range, % Agreement Denominator: 1.00, 1 to 1, 100%

Kappa, Range, % Agreement Numerator: 1.00, 1 to 1, 100%

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

Recommended guidelines for interpreting Kappa values from the National Quality Forum’s Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties
0 is "no better than chance"
0.01 to 0.20 is "slight"
0.21 to 0.40 is "fair"
0.41 to 0.60 is "moderate"
0.61 to 0.80 is "substantial"
0.81 to 1.0 is "almost perfect"

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 Tuberculosis Screening 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 | 8.5 | 0 | 14 | 14 | 0% |

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 1.0 for overall performance indicates high agreement between the automated report and the front-end chart abstraction.

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***](#section2b4)

**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***](#section2b5)***.***

**2b3.1. 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

**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***](#question2b49)

**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)*

Testing was performed in one large health system, which is a leader nationally in advancing patient safety. This health system has worked for over a decade to build systems to measure and improve quality of care in RA, and this is reflected in perfect performance (100%). There was therefore no statistical variation between providers at this site.

However, the proposed e-measure is analogous to the TB testing measure that has been part of the PQRS program since 2008. Data available from the ACR’s Rheumatology Clinical Registry suggest variation between providers with performance improving over time (*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*). Data from 2011 reveal performance of 73.6% among participating providers, increasing to 92.9% in 2012.

We also evaluated the variation in measure performance in 2017 among 105 RISE practices, representing 96.3% 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*)

National benchmarking data for this e-measure are currently not available. However, above we describe performance on the analogous PQRS measure as measured by the ACR’s Rheumatology Clinical Registry.

**Table 4. Variation in performance on Rheumatoid Arthritis: Tuberculosis Screening 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** |
| 105 | 9943 | 94.70 | 6-535 | 6074 | 57.85 | 0-324 | 58.85% | 43.75, 65.22, 78.15, 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?*)

See above.

The results demonstrate both wide variation and a continued need for improvement in performance overall given that the average performance in 2017 was 58.85%; 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 an extremely important standard of care required to protect patients from potential reactivation of TB. An average measure score under 60% (and a 75th percentile of 78%) 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*)

During the critical data element validity testing, missing data were encountered in 3% of patient records when testing this measure. In one instance, TB testing was performed but it was not recorded in a structured data field. This informed the ACR’s decision to move the measure designation from a true eCQM to a measure based-upon data abstracted from the EHR.

As noted above, the data abstraction approach ensures there is no missing data. See 2b6.3.

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

We expect that most practices will have a significant amount of missing data if similar workflows are not put into place. For example, it is not uncommon for TB testing to occur at a clinic or facility that differs from the clinic starting the biologic drug. In this case, TB test results may appear as scanned documents or free text in a clinical note. An automated report from the electronic record that draws information from structured fields such as laboratory results or immunizations will therefore underestimate performance.

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.