Arthritis Care & Research
Vol. 0, No. 0, Month 2019, pp 1–9
DOI 10.1002/acr.24040

American College of Rheumatology
Empowering Rheumatology Professionals

SPECIAL ARTICLE

2019 American College of Rheumatology Recommended Patient-Reported Functional Status Assessment Measures in Rheumatoid Arthritis

Claire E. H. Barber,[1] JoAnn Zell,[2] Jinoos Yazdany,[3] Aileen M. Davis,[4] Laura Cappelli,[5] Linda Ehrlich-Jones,[6] Donna Everix,[7] J. Carter Thorne,[8] Victoria Bohm,[1] Lisa Suter,[9] Alex Limanni,[10] and Kaleb Michaud[11]

Objective. To develop American College of Rheumatology (ACR) recommendations for patient-reported Functional Status Assessment Measures (FSAMs) for use in routine clinical practice in patients with rheumatoid arthritis (RA).

Methods. We convened a workgroup to conduct a systematic review of published literature through March 16, 2017 and abstract FSAM properties. Based upon initial search results and clinical input, we focused on the following FSAMs appropriate for routine clinical use: the Health Assessment Questionnaire (HAQ) and derived measures and the Patient-Reported Outcomes Measurement Information System (PROMIS) tool. We used the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) 4-point scoring method to evaluate each FSAM, allowing for overall level of evidence assessment. We identified FSAMs fulfilling a predefined minimum standard and, through a modified Delphi process, selected preferred FSAMs for regular use in most clinic settings.

Results. The search identified 11,835 articles, of which 56 were included in the review. Descriptions of the measures, properties, study quality, level of evidence, and feasibility were abstracted and scored. Following a modified Delphi process, 7 measures fulfilled the minimum standard for regular use in most clinic settings, and 3 measures were recommended: the PROMIS physical function 10-item short form (PROMIS PF10a), the HAQ-II, and the Multidimensional HAQ.

Conclusion. This work establishes ACR recommendations for preferred RA FSAMs for regular use in most clinic settings. These results will inform clinical practice and can support future ACR quality measure development as well as highlight ongoing research needs.

INTRODUCTION

Functional status is an important outcome in rheumatology and relates to measures of functioning that capture the interaction between a person's health condition and their ability to participate in activities (1). Poor functional status is associated with work disability (2), poor quality of life (3), and is one of the strongest predictors of mortality in rheumatoid arthritis (RA) (2,4–7). Functional

Supported by the American College of Rheumatology. Dr. Barber's work was supported by the Canadian Institutes of Health Research and the University of Calgary Department of Medicine. Dr. Davis' work was supported by the CIHR. Dr. Cappelli's work was supported by the Jerome Greene Foundation and Bloomberg Philanthropies. Dr. Ehrlich-Jones' work was supported by the National Institute on Disability, Independent Living, and Rehabilitation Research, the Craig H. Neilsen Foundation, the NIH (National Institute of Arthritis and Musculoskeletal and Skin Diseases grant 5R01-AR071091 and National Heart, Lung, and Blood Institute grant 1R01-HL132978-01A1), and the DOD. Dr. Thorne's work was supported by Mount Sinai and the University Health Network. Dr. Suter's work was supported by the VA and Yale New Haven Health System Center for Outcomes Research & Evaluation through a contract to the Centers for Medicare and Medicaid Services. Dr. Michaud's work was supported by the Rheumatology Research Foundation.

[1]Claire E. H. Barber, MD, PhD, FRCPC, Victoria Bohm, MSc, MPH: University of Calgary, Calgary, Alberta, Canada; [2]JoAnn Zell, MD: Denver Health, Denver, Colorado; [3]Jinoos Yazdany, MD, MPH: University of California, San Francisco; [4]Aileen M. Davis, PhD: Krembil Research Institute, University Health Network, and University of Toronto, Toronto, Ontario, Canada; [5]Laura Cappelli, MD: Johns Hopkins University, Baltimore, Maryland; [6]Linda Ehrlich-Jones, PhD, RN: Shirley Ryan AbilityLab, Chicago, Illinois; [7]Donna Everix, MPA, BS, PT: Mills Peninsula Health Services, Burlingame, California, and OnMyCare Home Health, Fremont, California; [8]Carter Thorne, MD, FRCPC: University of Toronto, Toronto, Ontario, Canada; [9]Lisa Suter, MD: Yale University, New Haven, Connecticut, and Veterans Affairs Medical Center, West Haven, Connecticut; [10]Alex Limanni, MD: Arthritis Centers of Texas, Dallas; [11]Kaleb Michaud, PhD: University of Nebraska Medical Center, Omaha, and FORWARD, the National Databank for Rheumatic Diseases, Wichita, Kansas.

Dr. Zell has received research support from Novartis. Dr. Yazdany has received consulting fees and/or speaking fees from Astra Zeneca and PRIME Education (less than $10,000 each) and research support from Pfizer and Astra Zeneca. Dr. Cappelli has received consulting fees from Regeneron/Sanofi (less than $10,000) and research support from Bristol-Myers Squibb. Dr. Ehrlich-Jones has received consulting fees from Zimmer Biomet (less than $10,000). Dr. Thorne has received consulting fees and/or speaking fees from AbbVie, Celgene, Novartis, Pfizer, Sandoz, Sanofi Genzyme, Serene, Amgen, and Medexus (less than $10,000 each). Dr. Limanni has received research support from GlaxoSmithKline, Janssen, Novartis, Gilead, and Pfizer. Dr. Michaud has received research support from Pfizer (ASPIRE award). No other disclosures relevant to this article were reported.

Address correspondence to Kaleb Michaud, PhD, Division of Rheumatology & Immunology, University of Nebraska Medical Center, 986270 Nebraska Medical Center, Omaha, NE 68198. E-mail: kmichaud@unmc.edu.

Submitted for publication April 17, 2019; accepted in revised form August 8, 2019.

status assessment measures (FSAMs) may be used in assessment of prognosis and aid in RA treatment decisions. Because of its importance, functional status assessment is included in guidelines for rheumatologic care for a number of conditions including RA (8). Assessment of functional status is captured by an American College of Rheumatology (ACR) RA quality measure (9) and is included in the Merit-Based Incentive Payment System, 1 of 2 payment tracks under the Quality Payment Program in the US emphasizing a value-based payment model (10).

In 2012 the ACR published recommendations on 6 RA disease activity measures (11). While no formalized document for ACR FSAM recommendations was developed, current ACR guidelines list collection of a standardized, validated FSAM as a key principle of RA treatment (8) and cite examples of commonly used FSAMs, including the Health Assessment Questionnaire (HAQ) disability index (DI), HAQ-II, Multidimensional HAQ (MDHAQ), and Patient-Reported Outcomes Measurement Information System (PROMIS) FSAMs, but do not make *specific* recommendations about their use in clinical practice. This work to provide initial recommendations on RA FSAMs was performed in parallel to an ACR workgroup updating the ACR's prior RA disease activity instrument recommendations.

The objectives of the RA FSAM workgroup were to provide RA patient-reported FSAMs meeting a minimum standard for regular use and preferred RA patient-reported FSAMs for regular use. These objectives reflect the fact that feasibility and clinical efficiency are important considerations in functional status assessment, supplementing minimum instrument performance standards.

METHODS

Study design. The ACR convened a workgroup of rheumatology professionals and rheumatologists to evaluate and recommend RA FSAMs. The workgroup developed a protocol and presented the process and preliminary findings at the 2017 ACR Annual Scientific Meeting in San Diego, California and obtained public comment following that presentation.

Search strategy. We conducted a systematic literature review, adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis checklist (12). We searched Medline, Embase, Cochrane Library, and the Cumulative Index of Nursing and Allied Health databases, from study inception to March 16, 2017. We devised search terms according to a published search strategy for finding studies on measurement properties of patient-reported outcome instruments (13) from the Consensus-Based Standards for the selection of Health Measurement Instruments (COSMIN) group (URL: http://www.cosmin.nl/). This strategy uses MeSH terms and keywords across 3 themes: construct search (for assessment of functional status), population search (RA), and instrument search (including terms for instruments of interest, e.g., questionnaires, etc.). The Boolean search operator "AND" was used to combine the 3 search themes (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). We manually searched the reference lists of included articles to identify potentially relevant studies. Additionally, we contacted content experts to ensure search completeness. We reviewed reference lists of relevant published reviews. Included articles were hand-searched for any additional relevant publications.

Eligibility criteria and article selection. We included studies with the primary objective of developing, validating, or establishing psychometric properties of patient-reported FSAMs in RA. We applied the following exclusion criteria: non-English publications, studies validating FSAMs in non-RA populations, performance-based measures (e.g., grip strength, walk tests, etc.), FSAMs that assessed a single extremity or body part, and studies using FSAMs to validate another instrument (e.g., assessing validity of joint ultrasound using FSAMs). We excluded health-related quality of life measures or multidimensional measures including function as a single construct among many (e.g., Short Form 36 [SF-36]) and studies only evaluating the cross-cultural validity of FSAMs.

Two reviewers (CEHB and JZ) first independently screened titles and abstracts to determine eligible studies for full text review and then conducted a full text review of eligible studies independently in duplicate. Disagreements between reviewers were resolved by discussion between reviewers or with a third reviewer (KM) when necessary.

Data abstraction and study quality assessment. Two of the 3 independent reviewers (CEHB, JZ, or VB) conducted data abstraction in duplicate for 15% of included articles to obtain consistent abstraction. A single reviewer (CEHB) abstracted the remaining studies with additional spot-checking of data abstraction performed by a second reviewer (VB). All measure characteristics, including details on measure items, administration time, scoring, and interpretation were abstracted. FSAMs with limited publications in RA (≤3) and/or not commonly used in the US (as evidenced in the ACR's Rheumatology Informatics System for Effectiveness [RISE] registry [14]) were not further evaluated for methodologic quality using COSMIN as it was unlikely such measures would be recommended for use due to feasibility concerns.

We rated the methodologic quality of included studies using COSMIN checklists (15). Briefly, COSMIN is a standardized tool for assessing study properties including internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, responsiveness, and interpretability. For each measurement property, a checklist of 5–18 items is completed and rated on a 4-point scale (poor, fair, good, or excellent) based on predefined criteria. An overall score for each property is based on the lowest score for each checklist. To assess the study psychometric result quality, we

employed a rating scheme using criteria proposed by Terwee et al (16) as modified by Dobson et al (17).

Although not rated using the 4-point scale, COSMIN reporting also includes standardized abstraction of items relating to the interpretability of the measurement property (including percentage of missing items and handling of missing items, adequate sample size, floor and ceiling effects, and minimum important change or minimum important difference) and the generalizability of the study (including population characteristics and study setting) (16).

Level of evidence. We provided the level of evidence for each individual FSAM psychometric property, considering all studies evaluating each property and their result using criteria proposed by Hendrikx et al (18) (Table 1). Each RA FSAM psychometric property received a level of evidence rating of strong (+++ or – – –), moderate (++ or – –), limited (+ or –), conflicting (±), or unknown (?) (Table 2). Three authors (CEHB, JZ, and VB) defined the level of evidence, with disagreements settled by a fourth author (KM).

Feasibility. Although evaluating the administration feasibility of FSAMs is not part of COSMIN, the workgroup agreed it is integral to making a recommendation for routine clinical use. An overall feasibility assessment for each FSAM was based on the following criteria: number of questions, whether computer-based administration was required, and associated costs or use licenses. The overall feasibility was scored as very feasible = +++, moderately feasible = ++, feasible = +, and not feasible = –.

Selection process. Ten workgroup members identified and selected by the ACR Quality Measures Subcommittee Chairs, including clinicians and researchers with expertise in functional status measurement and an ACR Quality Measures Subcommittee Liaison (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract) participated in a modified Delphi process to provide recommendations for the routine use of each FSAM. Only FSAMs with an overall assessment of adequate psychometric properties and feasibility (a rating of at least + on both) were reviewed. Members were given the study protocol and systematic review, including all COSMIN ratings and overall assessments. Prior to proceeding, members rated their comfort level with the study protocol and transparency, including the proposed modified Delphi process. During each of 3 rounds of the modified Delphi process, members rated each FSAM for ACR recommendation on a scale of 1 to 9 (where 1 = not recommended and 9 = essential to have). Following each round, members reviewed the results prior to re-rating. Following Round 2, workgroup members participated in a conference call to review and discuss the voting results, followed by a final round of voting. FSAMs were recommended if >80% of members (all but 1 member) rated the FSAM in the 7–9 range and excluded if >80% of ratings were in the 1–3 range, following best practices (19). FSAMs not achieving recommendation for inclusion or exclusion were deemed inconclusive. FSAMs deemed inconclusive at the end of voting remained on the list of measures fulfilling the minimum standard. The ACR Quality Measures Subcommittee reviewed these recommendations in parallel with the recommendations on functional status assessment, modifying as necessary based upon the goal of identifying preferred tools for regular use in most clinic settings, before voting. The ACR Quality of Care Committee and ACR Board of Directors reviewed and approved this article prior to publication.

Table 1. Rating the levels of evidence for the Functional Status Assessment Measures*

Level	Rating	Criteria
Strong	+++ or – – –	Consistent findings in multiple studies of good (methodologic) quality OR in one study of excellent quality
Moderate	++ or – –	Consistent findings in multiple studies of fair methodologic quality OR in one study of good methodologic quality
Limited	+ or –	One study of fair methodologic quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodologic quality
No evidence	0	No studies

* Positive result = +; negative result = −. Based on ref. 18.

RESULTS

A total of 11,835 articles underwent title and abstract screening; of those, 649 were eligible for full text review during which 571 articles were excluded (Figure 1). We identified 3 additional articles through hand searches, resulting in 81 included articles. After excluding 25 articles that were not based on the HAQ or PROMIS, 56 were subjected to COSMIN review, including 48 on HAQ-derived and 8 on PROMIS-derived instruments.

Patient-reported FSAMs. FSAMs ranged from simple visual analog scales to questionnaires with over 100 items (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). We excluded 19 FSAMs that had ≤3 RA-relevant publications and/or were rarely used in the US. The HAQ DI, 3 additional HAQ-derived measures (the modified HAQ [M-HAQ], MDHAQ, and HAQ-II), two PROMIS static forms (the physical function 10-item and 20-item [PF10a and PF20a]), and the PROMIS physical function Computer Adaptative Test (PF CAT) underwent COSMIN evaluation. Characteristics of included studies are shown in Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract.

Table 2. Overall assessment of the psychometric properties of the evaluated Functional Status Assessment Measures in rheumatoid arthritis*

	HAQ				PROMIS		
Psychometric properties	HAQ DI	M-HAQ	MDHAQ	HAQ-II	PF10a	PF20a	PF CAT
Internal consistency	++	++	++	++	0	0	++
Reliability							
Retest	++	?	?	0	0	+	+
Interrater	?	0	0	0	0	0	0
Measurement error	?	++	0	0	0	0	++
Validity							
Structural	+++	++	–	+	0	0	0
Criterion	N/A	++	0	+	0	0	N/A
Hypothesis testing	++	++	++	+	++	++	++
Content	+	0	0	0	0	+++†	0
Responsiveness‡	++	++	0	+	++	++	++
Interpretability	+/–	–	+	++	++	++	++
Overall assessment§	+	+	+	++	++	++	++

* HAQ = Health Assessment Questionnaire; HAQ DI = HAQ disability index; M-HAQ = modified HAQ; MDHAQ = Multidimensional HAQ; PROMIS = Patient Reported Outcomes Measurement Information system; PF10a = PROMIS physical function 10-item form; PF20a = PROMIS physical function 20-item form; PF CAT = PROMIS physical function Computer Adaptive Test.
† This study also examined content validity of the entire PROMIS item bank.
‡ Due to substantial heterogeneity in the evaluation of responsiveness, due to a lack of a functional status gold standard, only the quality of the studies was considered, not the result.
§ Overall assessment: + was assigned if the measures demonstrated adequate psychometric qualities (i.e., the measure is valid for use in routine clinical practice and captures functional status and can be reliably followed over time), ++ was assigned if, in addition, the measure had evidence of superior development methodology resulting in a more robust measure with improved floor/ceiling effects, and +++ was assigned if there was an abundance of evidence supporting a superiorly developed measure. Ratings of – were reserved for measures without any evidence of basic validity for use in routine clinical practice.

Internal consistency. There was moderate evidence for all HAQ-derived measures and the PROMIS PF CAT, which were the instruments with available internal consistency data (Table 2 and Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). Cronbach's alpha was the most commonly reported internal consistency assessment and was always acceptable (α = 0.70–0.95) when reported.

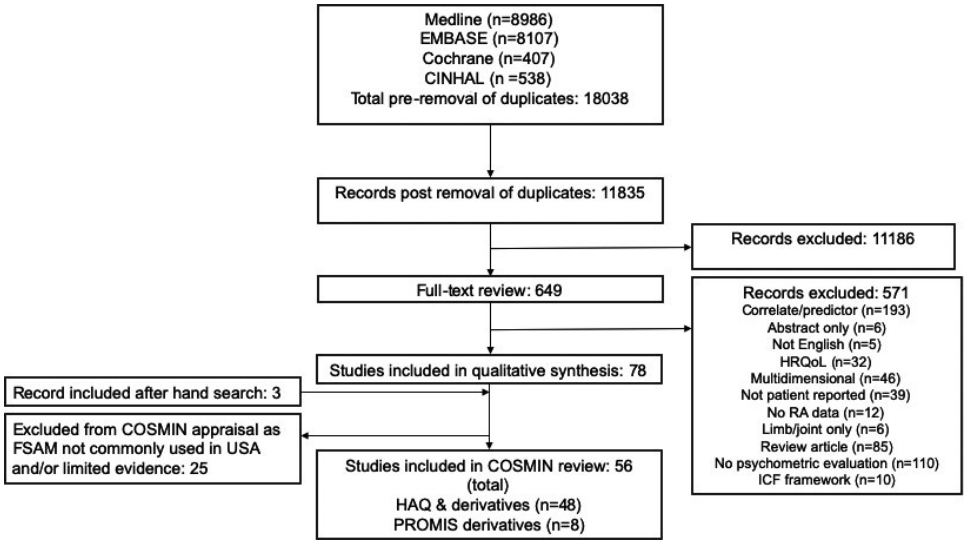

Figure 1. Flow diagram depicting manuscript selection for systematic review of functional status measures. COSMIN = Consensus-Based Standards for the Selection of Health Measurement Instruments; FSAM = Functional Status Assessment Measure; HAQ = Health Assessment Questionnaire; PROMIS = Patient-Reported Outcomes Measurement; HRQoL = health-related quality of life; RA = rheumatoid arthritis; ICF = International Classification of Functioning.

Reliability. The most common type of reliability testing, test–retest reliability, was usually assessed by interclass correlation coefficient (ICC). Reported ICCs were >0.7 for most domains (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). The HAQ DI reached a moderate reliability due to a single good COSMIN-rated study. Both the M-HAQ and MDHAQ had indeterminate reliability ratings because of only poor-quality studies. PROMIS measures had very limited reliability data and achieved a limited reliability rating for one FSAM.

Measurement error. According to COSMIN, the preferred measurement error statistics for classical test theory (CTT)–based studies are, in order of preference, standard error of measurement, limits of agreement, and smallest detectable change. Measurement error was only reported for the HAQ DI, M-HAQ, and PROMIS PF CAT, and each used a different method, which made comparisons challenging (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). The HAQ DI had only poor-quality studies, leading to an indeterminate assessment. The M-HAQ had a single fair-quality study that only provided 95% confidence intervals, supporting greater precision with an Item Response Theory (IRT)–based FSAM combining the SF-36 and M-HAQ than a non-IRT based measure (20). IRT-based measures use an item bank with specific questions related to a domain of health (21,22) that are evaluated for their correlation with a latent trait, in this case physical function (23). For the PROMIS PF CAT, study methods precluded COSMIN rating (24). However, results of the single study showed the PROMIS PF CAT had higher precision than the HAQ DI, based on root mean square errors. No study reported minimum important change, which should be greater than measurement error (16).

Content validity. The COSMIN content validity checklist assesses whether the authors appropriately judge item relevance and comprehensiveness. Very few articles explicitly evaluated RA FSAM content validity (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). A single, fair-quality article on the HAQ DI (25) yielded a limited rating. A study by Oude Voshaar et al (24) compared the PROMIS PF20, the PROMIS physical function item bank, the HAQ DI and the SF-36 physical function scale to the International Classification of Functioning, Disability, and Health (ICF) core set (26,27) for RA. Their high-quality study demonstrated that the PROMIS physical function item bank more comprehensively reflected all areas of RA-related physical function according to the ICF core set.

Structural validity. COSMIN structural validity reflects the "degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured" (i.e., functional status) (15). Factor analysis is the preferred CTT method, while IRT methods may also check item dimensionality. For good FSAM structural validity, factors should explain at least 50% of the variance (17). We identified 10 studies evaluating structural validity for the HAQ DI, M-HAQ, MDHAQ, and HAQ-II (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). Not all reported the percentage of variance explained by the models, because many used IRT-based methods, making comparisons challenging. In IRT, the model fit is examined to ensure the model reflects the true relationship between the underlying construct and the item response (28). Fit (or conversely misfit) of items describes the relationship between predicted and observed responses (28). One excellent study on the HAQ DI (29) yielded an overall strong weighting for structural validity despite lower-quality studies suggesting some misfitting HAQ items. We found 3 studies on M-HAQ (1 excellent, 1 fair, and 1 poor quality). However, the results of the methodologically strongest M-HAQ study concluded that an IRT-based scale combining the M-HAQ and SF-36 physical function scale had improved model fit versus the M-HAQ alone (20). The fair- and poor-quality studies identified misfitting M-HAQ items (2,30). A single, fair-quality HAQ-II study (2) demonstrated excellent structural validity compared to the HAQ DI, M-HAQ, and MDHAQ; however, limited evidence led to an overall low rating. The MDHAQ received a limited negative overall rating based upon 1 poor- (30) and 1 fair-quality study (2), which concluded the MDHAQ had 3 misfitting items. No study reported structural validity for the PROMIS-related measures in RA populations.

Criterion validity. Criterion validity assesses the degree to which instrument scores adequately reflect a gold standard. While there is no gold standard for RA FSAMs, in the case of HAQ-derived measures, the HAQ DI is considered the gold standard. Criterion validity evidence was assessed for the M-HAQ and HAQ-II (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). Given the fact that there were multiple studies of fair quality (2,31–33), the M-HAQ was assigned a moderate level of evidence. The HAQ-II received a limited evidence level based on a single fair-quality study (2).

Convergent validity. We found many instruments and variables assessing convergent validity between FSAMs, leading to heterogeneous results (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). Evidence of convergent validity was found for all instruments. However, the quality and number of studies varied, yielding a moderate level of evidence for all FSAMs except for HAQ-II. With only 1 fair-quality study, the HAQ-II received a limited rating (2).

Responsiveness. Responsiveness reflects an instrument's ability to detect change over time when true change has occurred. We identified responsiveness evidence for all FSAMs except the MDHAQ (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract). COSMIN stipulates that hypotheses about expected change scores or correlations between instrument change scores and changes in other variables should be expressed. Hypotheses about expected effect size or similar measures including standardized response means can also be used when explicit hypotheses are made. Heterogeneity in approach across studies made comparisons using our selected approach difficult. Furthermore, FSAM responsiveness testing used disparate comparator outcomes (e.g., patient's perception of change, pain, disease activity, etc.). Based only on study quality (and not the results due to significant reporting heterogeneity), we found moderate evidence for the HAQ DI, HAQ-II, M-HAQ, and all PROMIS measures.

Floor and ceiling effects. According to the results of a study by Terwee et al (16), fewer than 15% of respondents achieve the highest or lowest possible scores in good quality instruments. Where evaluated, the M-HAQ had high percentages of patients with the lowest scores leading to an unfavorable overall rating. There was mixed information about the HAQ. The HAQ-II, MDHAQ, and PROMIS measures achieved moderate ratings (see Supplementary Appendix A, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24040/abstract).

Results of feasibility. While the HAQ DI, M-HAQ, MDHAQ, HAQ-II, and the PROMIS measures are all feasible because they are in current use in clinical practice, the shorter FSAMs (the M-HAQ, MDHAQ, HAQ-II, and PROMIS PF10a) received higher feasibility ratings (Table 3). The PROMIS PF CAT received a lower rating due to computer and proprietary software requirements.

Delphi selection of recommended measures. The results from the modified Delphi process are shown in Table 4. The PROMIS PF10a and HAQ-II reached consensus for recommended use and no FSAMs reached consensus for exclusion. Among FSAMs without consensus, the M-HAQ had the lowest mean panelist score and the MDHAQ had the highest mean score (3.1 and 6.6, respectively).

The ACR Quality Measures Subcommittee approved these 2 recommendations with only 1 modification, which was the additional recommendation of the MDHAQ. The MDHAQ was included in the measures for preferred use based upon Delphi rating, feasibility, current use, and strength of its inclusion in the prior (11) and concurrent (34) ACR RA disease activity measure recommendations within the Routine Assessment of Patient Index Data 3 (RAPID3), considerations beyond this current work that focused solely on function.

DISCUSSION

This work represents the first ACR recommendations on FSAMs for use in routine clinical practice in RA. It provides a systematic literature review and synthesis of the psychometric properties of widely used FSAMs as well as a modified Delphi expert panel process to assess the feasibility of routine clinical use. Only 3 FSAMs were recommended: the PROMIS PF10a, HAQ-II, and MDHAQ. Consensus for recommendation was not reached for an additional 4 measures (the HAQ DI, M-HAQ, PROMIS PF20a, and PROMIS PF CAT). These 4 additional FSAMs will be monitored for inclusion in future recommendations along with any new instruments. Importantly, an inconclusive recommendation when applied to the 4 measures in this article should not necessarily prevent these measures from being used. Rather, it highlights the fact that more information is necessary before recommending widespread use of these 4 measures over other measures.

The HAQ DI (35) is one of the oldest and most widely-used patient-reported FSAMs in rheumatology. A variety of adaptations of the HAQ DI were later developed to shorten the

Table 3. Feasibility of the Functional Status Assessment Measures reviewed*

	HAQ				PROMIS		
Feasibility properties	HAQ DI	M-HAQ	MDHAQ	HAQ-II	PF10a	PF20a	PF CAT
No. of questions	20†	8	10	10	10	20	Variable (~5)
Requires computer	No	No	No	No	Assessment center scoring preferred‡	Assessment center scoring preferred‡	Yes§
Proprietary license for use	No	No	No	No	No	No	Yes
Overall feasibility assessment	++	+++	+++	+++	+++	++	+

* HAQ = Health Assessment Questionnaire; HAQ DI = HAQ disability index; M-HAQ = modified HAQ; MDHAQ = Multidimensional HAQ; PROMIS = Patient Reported Outcomes Measurement Information System; PF10a = PROMIS physical function 10-item form; PF20a = PROMIS physical function 20-item form; PF CAT = PROMIS physical function Computer Adaptive Test; +++ = very feasible; ++ = moderately feasible; + = feasible; – = not feasible.
† Requires assessment of the use of 13 assistive devices or help from others with 8 activities, and examined content validity of the entire PROMIS item bank.
‡ Score conversion tables available.
§ Assessment center pricing is available at URL: http://www.healthmeasures.net/resource-center/about-us/pricing-for-services.

Table 4. Results from 3-round modified Delphi process for functional status assessment measures*

	HAQ				PROMIS		
	HAQ DI	M-HAQ	MDHAQ	HAQ-II	PF10a	PF20a	PF CAT
Round 1							
Mean	6.4	5.3	5.1	6.9	7.1	6.5	5.6
Ratings†	0/6/4	3/3/4	3/4/3	1/1/8	1/0/9	1/2/7	1/5/3
Round 2							
Mean	6.4	3.6	4.4	7.1	N/A	6.6	5.3
Ratings†	1/3/6	6/3/1	5/1/4	1/0/9	N/A	1/1/8	2/6/2
Round 3							
Mean	6.2	3.1	6.6	N/A	N/A	6.5	5.7
Ratings†	1/4/5	6/4/0	0/3/7	N/A	N/A	1/2/7	3/1/6
Final recommendation	Inconclusive	Inconclusive	Recommended‡	Recommended	Recommended	Inconclusive	Inconclusive

* HAQ = Health Assessment Questionnaire; HAQ DI = HAQ disability index; M-HAQ = modified HAQ; MDHAQ = Multidimensional HAQ; PROMIS = Patient Reported Outcomes Measurement Information system; PF10a = PROMIS physical function 10-item form; PF20a = PROMIS physical function 20-item form; PF CAT = PROMIS physical function Computer Adaptive Test; N/A = not applicable because measure included based on previous rounds of voting.
† Ratings were reported by the number of participant votes on a 1–9 Likert scale (1–3/4–6/7–9) where 1–3 = not recommended, 4–6 = sometimes recommended, 7–9 = essential to have; and >80% agreement required for recommendation.
‡ During review by the American College of Rheumatology (ACR) Quality Measures Subcommittee, the additional final recommendation of the MDHAQ for preferred use was based upon Delphi rating, feasibility, current use, and strength of its inclusion in the prior and concurrent ACR Rheumatoid Arthritis disease activity measure recommendations within the Routine Assessment of Patient Index Data 3 measure.

scale while maintaining or improving its original psychometric properties. The most commonly used adaptations include the M-HAQ (32), MDHAQ (36), and the HAQ-II (2). More recently, PROMIS measures have been developed and are widely used (URL: http://www.nihpromis.org). PROMIS is a National Institutes of Health initiative that aims to create a more efficient and precise resource for patient outcome measurement when compared to existing legacy instruments for use in a wide variety of chronic disease conditions (21). PROMIS measures evaluate physical, mental, and social health across different chronic conditions (37) and general population health (21). Although most FSAMs were developed using CTT, the PROMIS measures were developed using modern IRT methods. PROMIS measures are available in static short forms with a fixed number of questions and also as computer adaptive tests, which adapt to the ability level of the respondent. The results of all PROMIS measures are normalized to the US population and reported with a T score (mean ± SD 50 ± 10).

The PROMIS physical function measures evaluated in our study included the 10- and 20-item static forms (the PF10a and PF20a) and the PROMIS PF CAT. However, only the PROMIS PF10a was recommended by our panelists. While the PROMIS physical function measures were developed using rigorous methods and tested extensively in the general population and populations with chronic disease (22,38,39), there were few studies specific to patients with RA (24,40–45), impacting panelist ratings. Panelists concluded that the shorter 10-item instrument was likely more feasible for routine use in the clinic than the 20-item survey. While the adaptive PROMIS PF CAT usually requires the fewest items, the computer and proprietary software requirements reduced its feasibility.

The HAQ-II is a 10-item questionnaire developed using Rasch analysis and IRT-based methodology. Instrument development was aimed at addressing 4 main issues identified with the original HAQ DI and its derivatives: removing misfitting items, maximizing scale length, eliminating items with overlapping difficulties, and eliminating gaps in measurement along the continuum of functional status assessment (2). The resulting instrument includes 5 items from the original HAQ DI plus 5 new items. When compared to the M-HAQ, MDHAQ, and HAQ DI, the HAQ-II better captures the disability continuum. Gaps in the measurement of disability were found in all scales evaluated except the HAQ-II, indicating that the HAQ-II has the most favorable psychometric properties of the HAQ-derived instruments. The HAQ-II also has the least floor effect among the evaluated HAQ-derived measures.

Although the HAQ DI is the legacy FSAM, and has been extensively tested and used worldwide, its psychometric properties when compared to the HAQ-II and the newer PROMIS measures were felt to be less favorable. Additionally, the length and relatively complex scoring of the HAQ DI led to lower panelist ratings.

The MDHAQ was designed as a shorter version of the HAQ DI and includes 10 items (all items from the M-HAQ plus 2 additional items) (32). While the MDHAQ has greater feasibility than the original HAQ DI and more favorable psychometric properties compared to the M-HAQ (36), it performs less well when compared to the HAQ-II (2) or the PROMIS measures (44). A limitation in our assessment of the MDHAQ is that we did not evaluate the literature on the RAPID3 measure (46). The RAPID3 is a patient-reported disease activity tool that includes the MDHAQ, a measure of pain, and a patient global score (46). The psychometric and clinometric properties of the RAPID3 have been reviewed by the ACR RA Disease Activity Workgroup, which recommended the RAPID3 as an effective measure of RA disease activity. RAPID3 is also the most commonly collected disease activity measure in the RISE registry

(14). Given this, we additionally recommend the MDHAQ as a preferred FSAM.

The 8-item M-HAQ is derived from the HAQ DI (using 1 question from each domain) and is the shortest measure evaluated (22). Although the M-HAQ is highly correlated to the HAQ DI (32), the M-HAQ has significant floor effects and may not be as sensitive to clinical changes as longer scales (2). The panel did not reach consensus for excluding the M-HAQ; however, it received the lowest scores of all the FSAMs evaluated.

Our study had a number of strengths, including the rigorous and transparent methodologic assessment of the measures combined with expert opinion; however, there are some limitations. We did not subject all FSAMs to COSMIN assessment and consideration by our expert panel because it was felt unlikely that measures not already commonly used in the US would be included in our final recommendations. Therefore, it is possible that measures with highly favorable psychometric properties were not considered in generating our recommendations. Additionally, our review was conducted while only considering RA-specific data and English-language publications, and it is possible this limited the evidence on which our recommendations were based. After our systematic review was completed, the COSMIN group updated their checklist (47), and the study ratings could be different if the updated checklist was used. Given that the overall panelist ratings on the FSAMs weighed not only the psychometric properties as evaluated by COSMIN but also measured feasibility, it is less likely that the overall outcome of the process would have varied greatly from our present results by using the updated checklist. Patients were not involved in the panel, given the significant methodologic expertise required for the project; however, this work will inform ongoing measure development work, which includes patient partners. Lastly, given the paucity of psychometric data on some measures, further research in this area is warranted and it is possible that some of the recommendations may change in the future as a result of new findings.

In conclusion, we have presented the first ACR recommendations on FSAMs for routine use in clinical practice to be used for the assessment of functional status in RA, based on a rigorous systematic review and expert panel process. Although we only recommend 3 FSAMs, this work should not preclude the use of other identified measures but rather encourage the use of measures with the most favorable psychometric properties while highlighting the need for ongoing research in this area.

ACKNOWLEDGMENTS

We thank American College of Rheumatology staff members Amy Turner and Regina Parker for their support and assistance through the recommendation process.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Michaud had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Barber, Zell, Yazdany, Davis, Cappelli, Suter, Limanni, Michaud.

Acquisition of data. Barber, Zell, Bohm.

Analysis and interpretation of data. Barber, Zell, Davis, Cappelli, Ehrlich-Jones, Everix, Thorne, Suter, Michaud.

REFERENCES

1. World Health Organization. International classification of functioning, disability and health (ICF). May 2001. URL: https://www.who.int/classifications/icf/en/.
2. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.
3. Cohen JD, Dougados M, Goupille P, Cantagrel A, Meyer O, Sibilia J, et al. Health assessment questionnaire score is the best predictor of 5-year quality of life in early rheumatoid arthritis. J Rheumatol 2006;33:1936–41.
4. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. Ann Intern Med 1994;120:26–34.
5. Michaud K, Vera-Llonch M, Oster G. Mortality risk by functional status and health-related quality of life in patients with rheumatoid arthritis. J Rheumatol 2012;39:54–9.
6. Sokka T, Pincus T. Poor physical function, pain and limited exercise: risk factors for premature mortality in the range of smoking or hypertension, identified on a simple patient self-report questionnaire for usual care. BMJ Open 2011;1:e000070.
7. Yelin E, Trupin L, Wong B, Rush S. The impact of functional status and change in functional status on mortality over 18 years among persons with rheumatoid arthritis. J Rheumatol 2002;29:1851–7.
8. Singh JA, Saag KG, Bridges SL Jr, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol 2016;68:1–26.
9. Yazdany J, Robbins M, Schmajuk G, Desai S, Lacaille D, Neogi T, et al. Development of the American College of Rheumatology's rheumatoid arthritis electronic clinical quality measures. Arthritis Care Res (Hoboken) 2016;68:1579–90.
10. Department of Health & Human Services. Quality payment program: quality measures requirements. URL: https://qpp.cms.gov/mips/quality-measures.
11. Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, et al. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken) 2012;64:640–7.
12. PRISMA. Transparent reporting of systematic reviews and meta-analyses. URL: http://www.prisma-statement.org/.
13. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115–23.
14. Yazdany J, Bansback N, Clowse M, Collier D, Law K, Liao KP, et al. Rheumatology Informatics System for Effectiveness: a national informatics-enabled registry for quality improvement. Arthritis Care Res (Hoboken) 2016;68:1866–73.
15. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of

studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.

16. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
17. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage 2012;20:1548–62.
18. Hendrikx J, de Jonge MJ, Fransen J, Kievit W, van Riel PL. Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis. RMD Open 2016;2:e000202.
19. Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014;67:401–9.
20. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, MacLean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647–60.
21. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol 2010;63:1179–94.
22. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE Jr. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67:516–26.
23. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: assessment of physical function. J Rheumatol 2014;41:153–8.
24. Oude Voshaar MA, ten Klooster PM, Glas CA, Vonkeman HE, Taal E, Krishnan E, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity-driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology (Oxford) 2015;54:2221–9.
25. Goeppinger J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. Res Nurs Health 1988;11:321–31.
26. Cieza A, Brockow T, Ewert T, Amman E, Kollerits B, Chatterji S, et al. Linking health-status measurements to the international classification of functioning, disability and health. J Rehabil Med 2002;34:205–10.
27. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lessons learned. J Rehabil Med 2005;37:212–8.
28. Siemons L, Krishnan E. A short tutorial on item response theory in rheumatology. Clin Exp Rheumatol 2014;32:581–6.
29. Cole JC, Motivala SJ, Khanna D, Lee JY, Paulus HE, Irwin MR. Validation of single-factor structure and scoring protocol for the Health Assessment Questionnaire-Disability Index. Arthritis Rheum 2005;53:536–42.
30. Pincus T, Sokka T, Kautiainen H. Further development of a physical function scale on a Multidimensional Health Assessment Questionnaire for standard care of patients with rheumatic diseases. J Rheumatol 2005;32:1432–9.
31. Uhlig T, Haavardsholm EA, Kvien TK. Comparison of the Health Assessment Questionnaire (HAQ) and the modified HAQ (M-HAQ) in patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:454–8.
32. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment questionnaire. Arthritis Rheum 1983;26:1346–53.
33. Nagasawa H, Kameda H, Sekiguchi N, Amano K, Takeuchi T. Differences between the Health Assessment Questionnaire Disability Index (HAQ DI) and the modified HAQ (M-HAQ) score before and after infliximab treatment in patients with rheumatoid arthritis. Mod Rheumatol 2010;20:337–42.
34. England BR, Tiong BK, Bergman MJ, Curtis JR, Kazi S, Mikuls TR, et al. 2019 update of the American College of Rheumatology Recommended Rheumatoid Arthritis Disease Activity Measures. Arthritis Care Res [Hoboken] 2019. https://doi.org/10.1002/acr.24042. E-pub ahead of print.
35. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
36. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42:2220–30.
37. Cook KF, Jensen SE, Schalet BD, Beaumont JL, Amtmann D, Czajkowski S, et al. PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. J Clin Epidemiol 2016;73:89–102.
38. Hays RD, Spritzer KL, Amtmann D, Lai JS, Dewitt EM, Rothrock N, et al. Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult physical functioning item bank. Arch Phys Med Rehabil 2013;94:2291–6.
39. Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2010;63:1195–204.
40. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol 2011;38:1759–64.
41. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther 2011;13:R147.
42. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis 2015;74:104–7.
43. Oude Voshaar MA, ten Klooster PM, Glas CA, Vonkeman HE, Krishnan E, van de Laar MA. Relative performance of commonly used physical function questionnaires in rheumatoid arthritis and a patient-reported outcomes measurement information system computerized adaptive test. Arthritis Rheumatol 2014;66:2900–8.
44. Bartlett SJ, Orbai AM, Duncan T, DeLeon E, Ruffing V, Clegg-Smith K, et al. Reliability and validity of selected PROMIS measures in people with rheumatoid arthritis. PLoS ONE 2015;10:e0138543.
45. Wahl E, Gross A, Chernitskiy V, Trupin L, Gensler L, Chaganti K, et al. Validity and responsiveness of a 10-item patient-reported measure of physical function in a rheumatoid arthritis clinic population. Arthritis Care Res (Hoboken) 2017;69:338–46.
46. Pincus T, Swearingen CJ, Bergman M, Yazici Y. RAPID3 (Routine Assessment of Patient Index Data 3), a rheumatoid arthritis index without formal joint counts for routine care: proposed severity categories compared to disease activity score and clinical disease activity index categories. J Rheumatol 2008;35:2136–47.
47. Mokkink LB, de Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1171–9.