Title

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

Journal

Arthritis Care & Research, 71(12)

ISSN

2151-464X

Authors

Barber, Claire EH
Zell, JoAnn
Yazdany, Jinoos
et al.

Publication Date

2019-12-01

DOI

10.1002/acr.24040

Peer reviewed

HHS Public Access

Author manuscript

Arthritis Care Res (Hoboken). Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

Arthritis Care Res (Hoboken). 2019 December ; 71(12): 1531–1539. doi:10.1002/acr.24040.

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

Claire Barber, MD, PhD, FRCPC[1], **JoAnn Zell, MD**[2], **Jinoos Yazdany, MD, MPH**[3], **Aileen M. Davis, PhD**[4,8], **Laura Cappelli, MD**[5], **Linda Ehrlich-Jones, PhD, RN**[6], **Donna Everix, MPA, BS, PT**[7], **Carter Thorne, MD, FRCPC**[8], **Victoria Bohm**[1], **Lisa Suter, MD**[9], **Alex Limanni, MD**[10], **Kaleb Michaud, PhD**[11]

[1.]University of Calgary, Calgary, AB

[2.]Denver Health

[3.]Division of Rheumatology, University of California San Francisco, San Francisco, CA

[4.]Krembil Research Institute, University Health Network, Toronto, Canada

[5.]Johns Hopkins University, Baltimore, MD

[6.]Shirley Ryan AbilityLab

[7.]Mills Peninsula Health Services

[8.]University of Toronto, Toronto, ON

[9.]Yale University, New Haven, CT; Veterans Affairs Medical Center, West Haven, CT

[10.]Arthritis Centers of Texas, Dallas, TX

[11.]University of Nebraska Medical Center, Omaha, NE; FORWARD, The National Databank for Rheumatic Diseases, Wichita, KS

Abstract

Objectives: To develop 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 FSAMs appropriate for routine clinical use: the Health Assessment

Corresponding author: Kaleb Michaud, PhD. Associate Professor, Division of Rheumatology & Immunology, University of Nebraska Medical Center. 986270 Nebraska Medical Center, Omaha, NE 68198-6270. Phone: 402-559-7288. Fax: 402-559-6788. kmichaud@unmc.edu.

Disclosures: CB, JZ, AMD, DE, VB, AL, KM none. JY receives consultant fees from Astra Zeneca and speaker fees from PRIME Education. LC receives advisory board fees from Regeneron/Sanofi. LEJ receives consultant fees from American Insitute of Biological Sciences, Switzer Funding, Zimmer Biomet, Partners Healthcare, Computer Sciences Govt. Sols, University of Illinois at Chicago and Northwestern University. CT receives advisory board fees from Abbvie, Celgene, Novartis, Pfizer, Sandoz, Sanofi Genzyme and Serene, consultant fees from Amgen and speaker fees from Medexus. LGS develops and implements non-rheumatology accountability outcome measures under contract to the Centers for Medicare and Medicaid Services.

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 pre-defined 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 three measures were recommended: Patient-Reported Outcomes Measurement Information System physical function 10a form (PROMIS PF10a), Health Assessment Questionnaire-II (HAQ-II), and the Multidimensional Health Assessment Questionnaire (MD-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 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 included in the Merit-based Incentive Payment System, one of two payment tracks under the Quality Payment Program in the United States emphasizing a value-based payment model (10).

In 2012 the ACR published recommendations on 6 RA disease activity measures (RADAMs) (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 HAQ-DI, HAQ-II, MD-HAQ and 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 working group updating the ACR's prior RA disease activity instrument recommendations.

The objectives of the RA FSAM workgroup were to provide 1) RA patient-reported FSAMs meeting a minimum standard for regular use and 2) "preferred" RA patient-reported FSAMs for regular use. These objectives reflect 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 Meeting (San Diego, CA) and obtained public comment.

Search strategy

We conducted a systematic literature review, adhering to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) checklist (12). We searched MEDLINE, EMBASE, Cochrane Library and CINHAL databases, from 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 (http://www.cosmin.nl/). This strategy uses MeSH terms and keywords across three themes: #1 construct search (for assessment of functional status), #2 population search (rheumatoid arthritis) and #3 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 (Appendix Figure 1). 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 limb or body part, 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 (SF)-36) and studies only evaluating the cross-cultural validity of FSAMs.

Two reviewers (CB 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 3 independent reviewers (CB, JZ and VB) conducted data abstraction in duplicate for 15% of included articles to obtain consistent abstraction using a single reviewer (CB) who abstracted the remaining studies with additional spot-checking of data-abstraction by a second reviewer (VB). We abstracted all measure characteristics, including details on

measure items, administration time, scoring and interpretation. FSAMs with limited publications in RA (3 or fewer) and/or not commonly in use 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 four-point scale (poor, fair, good or excellent) based on pre-defined 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 four-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 minimally important change) and the generalizability of the study (including population characteristics and study setting) (16).

Level of Evidence

We provided level of evidence for each individual FSAM psychometric property, considering all studies evaluating each property and their result using criteria by Hendrix et al. (18) (Table 1). Each RA FSAM psychometric property received a level of evidence of: Strong (+++ or −−−), Moderate (++ or −−), Limited (+ or −), Conflicting (±), or Unknown (?) (Table 2). Three authors (CB, JZ, VB) defined the level of evidence, with disagreements settled by a fourth author (KM).

Feasibility

Although administration feasibility of FSAMs is not part of COSMIN, the workgroup agreed it is integral to making a recommendation for routine clinical use. Aan overall feasibility assessment for each FSAM was based on the following criteria: number of questions, whether computer-based administration is required, and associated costs or use licenses. We defined the overall feasibility as "+++" =very feasible, "++" =moderately feasible, "+" =feasible, "-" =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 (Appendix), 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 (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, members rated each FSAM for ACR recommendation (1=not recommended, to 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) 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 Quality of Care Committee and ACR Board reviewed and approved this manuscript prior to publication.

RESULTS

A total of 11835 articles underwent title and abstract screening; 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 not based on HAQ or PROMIS, 56 were subjected to COSMIN review, including 48 on HAQ-derived and 8 on PROMIS-derived instruments.

Patient-reported Functional Status Assessment Measures

FSAMs ranged from simple visual analogue scales (VAS) to questionnaires with over 100 items (Appendix Table 1). We excluded 19 FSAMs with 3 or fewer RA-relevant publications and/or rare US usage. The HAQ-DI, 3 additional HAQ-derived measures (MHAQ, MD-HAQ and HAQ-II), two PROMIS static forms (PF10a and PF20a) and the PROMIS PF CAT underwent COSMIN evaluation. Characteristics of included studies are shown in Appendix Table 2.

Internal consistency

We found moderate evidence for all HAQ-derived measures and the PROMIS PF CAT, which were the instruments with internal consistency data (Table 2, Appendix Table 3). Cronbach's alpha was the most commonly reported internal consistency assessment and was always acceptable (0.70–0.95) when reported.

Reliability

The most common type of reliability testing was test-retest reliability, usually assessed by interclass correlation (ICC). Reported ICCs were > 0.7 for most domains (Appendix Table 3). The HAQ-DI reached a moderate reliability due to a single "good" COSMIN-rated study. Both the M-HAQ and MD-HAQ had indeterminate reliability ratings as we identified 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 or preference, standard error of measurement (SEM), limits of agreement (LoA) and smallest detectable change. Measurement error was only reported for HAQ-DI, M-HAQ and PROMIS PF CAT and each used a different method making comparisons challenging (Appendix Table 3). HAQ-DI had only poor-quality studies, leading to an indeterminate assessment. M-HAQ had a single fair study that only provided 95% confidence intervals supporting greater precision with an Item Response Theory (IRT)-based FSAM combining 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, the single study concluded PROMIS PF CAT had higher precision than HAQ-DI, based on root mean square errors. No study reported minimal important change (MIC), 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 (Appendix Table 4). A single, fair quality article on the HAQ-DI (25) yielded a limited rating. Oude Voshaar et al. (24) compared the PROMIS PF20, PROMIS physical function item bank, HAQ-DI and SF-36 PF scale to the International Classification of Functioning, Disability and Health (ICF) core set (26, 27) for RA. Their high-quality study concluded the PROMIS physical function item bank most 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 HAQ-DI, M-HAQ, MD-HAQ and HAQ-II (Appendix Table 4). Not all reported the percentage of variance explained by the models, as 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 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 MHAQ (one excellent quality, one fair and one poor). However, the methodologically-strongest MHAQ study concluded that an IRT-based scale combining MHAQ and SF-36 PF scale had improved model fit versus MHAQ alone (20). The fair and

poor-quality studies identified misfitting MHAQ items (2, 30). A single, fair quality HAQ-II study (2) demonstrated excellent structural validity compared to HAQ-DI, MHAQ and MD-HAQ; however, limited evidence led to an overall low rating. MD-HAQ received a limited negative overall rating based upon one poor (30) and one fair quality study (2), which concluded MD-HAQ 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, for HAQ-derived measures, HAQ-DI was considered the "gold-standard". Criterion validity evidence was assessed for MHAQ and HAQ-II (Appendix Table 4). Given multiple studies of fair quality (2, 31–33), MHAQ was assigned a moderate level of evidence. 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 (Appendix Table 4). 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 with the exception of HAQ-II. With only one fair quality study, 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 MD-HAQ (Appendix Table 4). 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 (SRMs) 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 HAQ-DI, HAQ-II, MHAQ and all PROMIS measures.

Floor and Ceiling Effects

According to Terwee et al. (16), fewer than 15% of respondents achieve the highest or lowest possible scores in good quality instruments. Where evaluated, MHAQ had high percentages of patients with the lowest scores leading to an unfavorable overall rating. There was mixed information about HAQ. HAQ-II, MD-HAQ and PROMIS measures achieved moderate ratings (Appendix Table 4).

Feasibility

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

Delphi selection of recommended measures

Table 4 lists results from the modified Delphi process. PROMIS PF10a and HAQ-II reached consensus for recommended use and no FSAMs reached consensus for exclusion. Among FSAMs without consensus, MHAQ had the lowest mean panelist score (3.1) and MD-HAQ had the highest (6.6).

The ACR Quality Measures Subcommittee approved these two recommendations with a lone modification - the additional recommendation of MD-HAQ 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 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 asses feasibility of routine clinical use. Only three FSAMs are recommended: PROMIS PF10a, HAQ-II, and MD-HAQ. Consensus for recommendation was not reached for an additional four measures (HAQ-DI, MHAQ, PROMIS PF20a and PROMIS PF CAT). These FSAMs will be monitored for inclusion in future recommendations along with any new instruments. Importantly, “inconclusive” recommendations when applied in this paper should not necessarily prevent these four measures from being used; however, highlights that more information is necessary before recommending widespread use over other measures.

The Health Assessment Questionnaire Disability Index (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 scale while maintaining or improving its original psychometric properties. The most commonly used adaptations include the Modified HAQ (MHAQ) (32), the Multidimensional HAQ (MD-HAQ) (36), and the HAQ-II (2). More recently, “Patient-Reported Outcomes Measurement Information System” or PROMIS measures have been developed and are widely used (www.nihpromis.org). PROMIS is a National Institutes of Health (NIH) initiative to create a more efficient and precise resource for patient outcome measurement than existing legacy instruments for 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). While 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 (CATs), 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 of 50 and a standard deviation of 10).

The PROMIS physical function measures evaluated in our study included the 10 and 20-item static forms (PF-10a, 20a) and the PROMIS PF CAT. However, only one of these was recommended by our panelists, PROMIS PF10a. 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 are few studies specific to patients with RA (24, 40–45), impacting panelist ratings. Panelists concluded that the shorter 10-item instrument is likely more feasible for routine use in 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.

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 and 5 new items. When compared to the MHAQ, MD-HAQ 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 HAQ-II has the most favorable psychometric properties of HAQ-derived instruments. HAQ-II also has the least floor effect among evaluated HAQ-derived measures.

While HAQ-DI is the legacy FSAM, extensively tested and used worldwide, the 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.

MD-HAQ was designed as a shorter version of the HAQ-DI and includes 10 items including all items from the M-HAQ and 2 additional items (32). While the MD-HAQ has greater feasibility than the original HAQ-DI and more favorable psychometric properties compare to the MHAQ (36), it performs less well when compared to the HAQ-II (2) or the PROMIS measures (44). A limitation in our assessment of the MD-HAQ is that we did not evaluate the literature on the Routine Assessment of Patient Index Data 3 (RAPID3) measure (46). The RAPID3 is a patient-reported disease activity tool that includes the MD-HAQ, a measure of pain and also 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 MD-HAQ as a preferred FSAM.

Derived from HAQ-DI using one question from each domain, the 8-item MHAQ was the shortest measure evaluated (32). While MHAQ is highly correlated to HAQ-DI (32), it has significant floor effects and may not be as sensitive to clinical changes as longer scales (2).

Although the panel did not reach consensus for excluding the MHAQ, it had the lowest scores by far of the FSAMs evaluated.

While our study has a number of strengths, including the rigorous and transparent methodological assessment of the measures combined with expert opinion, there are some limitations. We did not subject all FSAMs to COSMIN assessment and consideration by our expert panel as it was felt unlikely that measures not already in common use in the US would be included in our final recommendations. Therefore, it is possible 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 upon which our recommendations were based. After our systematic review was completed, the COSMIN group updated their checklist (47) and study ratings could be different with the updated checklist. Given that the overall panelists ratings on the FSAMs weighed not only the psychometric properties as evaluated by COSMIN but also measure 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 future as a result of new findings.

In conclusion, we present the first ACR recommendations on FSAMs for routine use in clinical practice for the assessment of functional status in RA based on a rigorous systematic review and expert panel process. While we only recommend three 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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

Funding: Internal funding through the American College of Rheumatology

Other Support: CB is supported by funding from CIHR and the University of Calgary Department of Medicine. JZ is supported by funding from Novartis. JY is supported by funding from NIH/NIAMS, AHRQ, CDC, Pfizer and Astra Zeneca. AMD is supported by funding from CIHR. LC is supported by funding from Bristol-Meyers Squibb, the Jerome Greene Foundation and Bloomberg Philanthropies. LEJ is supported by funding from NIDILRR, the Craig H. Neilsen Foundation, NIH/NIAMS, Department of Defense and NIH/NHLBI. CT is supported by Mount Sinai and University Health Network. LGS receives salary support from the Veterans Administration and YNHHS/CORE through a contract to the Centers for Medicare and Medicaid Services to develop and implement accountability outcome measures. AL is supported by Glaxo, Janssen, Novartis, Gilead and Pfizer. KM was supported by funding from Pfizer ASPIRE award and RRF.

References

1. World Health Organization (WHO). World Health Organization: International Classification of Functioning, Disability and Health (ICF),. 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. [PubMed: 15476213]
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. [PubMed: 16924692]
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. [PubMed: 8250453]
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. [PubMed: 22089466]
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. [PubMed: 12233878]
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. [PubMed: 27564778]
10. Department of Health & Human Services USA. Quality Payment Program: Quality Measures. 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. [PubMed: 22473918]
12. PRISMA. PRISMA, Transparent reporting of systematic reviews and meta-analyses. 2015 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. [PubMed: 19711195]
14. Yazdany J, Bansback N, Clowse M, Collier D, Law K, Liao KP, et al. The Rheumatology Informatics System for Effectiveness (RISE): A National Informatics-Enabled Registry for Quality Improvement. Arthritis Care Res (Hoboken). 2016.
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. [PubMed: 21732199]
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. [PubMed: 17161752]
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. [PubMed: 22944525]
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. [PubMed: 27651921]

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. [PubMed: 24581294]

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. [PubMed: 17334829]

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. [PubMed: 20685078]

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. [PubMed: 24698295]

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. [PubMed: 24241485]

24. Oude Voshaar MAH, ten Klooster PM, Glas CAW, 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. 2015;54:2221–9. [PubMed: 26224306]

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. [PubMed: 3175056]

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. [PubMed: 12392234]

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. [PubMed: 16024476]

28. Siemons L, Krishnan E. A short tutorial on item response theory in rheumatology. Clin Exp Rheumatol. 2014;32:581–6. [PubMed: 25065775]

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. [PubMed: 16082630]

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. [PubMed: 16078316]

31. Uhlig T, Haavardsholm EA, Kvien TK. Comparison of the Health Assessment Questionnaire (HAQ) and the modified HAQ (MHAQ) in patients with rheumatoid arthritis. Rheumatology. 2006;45:454–8. [PubMed: 16287925]

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. [PubMed: 6639693]

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 (mHAQ) score before and after infliximab treatment in patients with rheumatoid arthritis. Mod Rheumatol. 2010;20:337–42. [PubMed: 20225006]

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. 2019.

35. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23:137–45. [PubMed: 7362664]

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. [PubMed: 10524697]

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. [PubMed: 26952842]
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. [PubMed: 23751290]
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. [PubMed: 20688471]
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. [PubMed: 21807798]
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. [PubMed: 21914216]
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. [PubMed: 24095937]
43. Oude Voshaar MAH, Ten Klooster PM, Glas CAW, Vonkeman HE, Krishnan E, Van De Laar MAFJ. 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. [PubMed: 24964773]
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. [PubMed: 26379233]
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. [PubMed: 27332620]
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. [PubMed: 18793006]
47. Mokkink LB, de Vet HCW, Prinsen CAC, 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. [PubMed: 29260445]

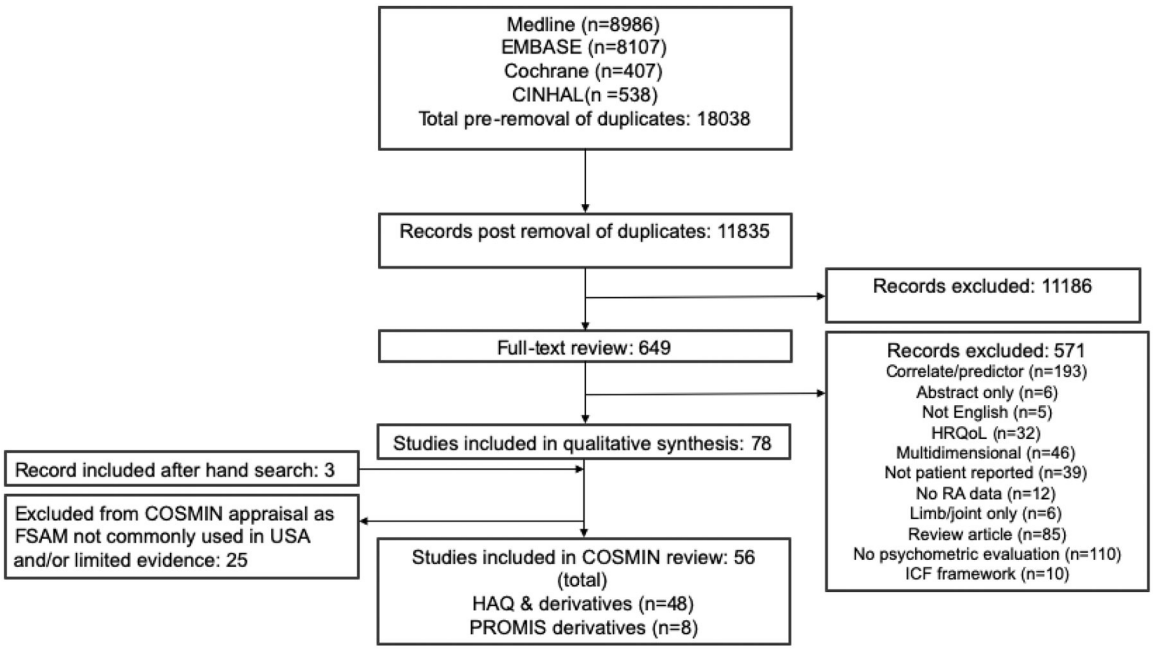

Figure 1.

Flow diagram depicting manuscript selection for systematic review of functional status measures

Abbreviations: Consensus-based Standards for the selection of health Measurement Instruments (COSMIN), Functional Status Assessment Measure (FSAM), Health Assessment Questionnaire (HAQ), Health-related quality of life (HRQoL), International Classification of Functioning (ICF), Patient-Reported Outcomes Measurement (PROMIS), Rheumatoid Arthritis (RA)

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 (methodological) quality OR in one study of excellent quality
Moderate	++ or − −	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or −	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality
No evidence	0	No studies

+ positive result, − negative result (Based on Hendrix et al. RMD Open 2016)

Table 2.

Overall Assessment of the Psychometric Properties of the Evaluated Functional Status Assessment Measures in Rheumatoid Arthritis

	HAQ				PROMIS		
	HAQ-DI	**MHAQ**	**MD-HAQ**	**HAQ-II**	**PF 10a**	**PF 20a**	**PF CAT**
Psychometric Properties							
Internal consistency	++	++	++	++	0	0	++
Reliability							
Retest	++	?	?	0	0	+	+
Inter-rater	?	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	+++[1]	0
Responsiveness[2]	++	++	0	+	++	++	++
Interpretability	+/−	−	+	++	++	++	++
Overall Assessment	+	+	+	++	++	++	++

[1] This study examined content validity of the entire PROMIS item bank as well

[2] 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 considered, not the result.

[3] Overall assessment: “+” was assigned if the measures demonstrated adequate psychometric qualities (i.e. the measure is valid for routine use in clinic 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

Table 3.

Feasibility of Functional Status Assessment Measures Reviewed

	HAQ				PROMIS		
	HAQ-DI	**MHAQ**	**MD-HAQ**	**HAQ-II**	**PF 10a**	**PF 20a**	**PF CAT**
Feasibility Properties							
Number of Questions	20[1]	8	10	10	10	20	Variable (~5)
Requires Computer	No	No	No	No	Assessment center scoring preferred[2]	Assessment center scoring preferred[2]	Yes[3]
Proprietary license for use	No	No	No	No	No	No	Yes
Overall Feasibility Assessment	++	+++	+++	+++	+++	++	+

CAT= Computer Adaptive Test

+++ = very feasible, ++ =moderately feasible, += feasible, − =not feasible

[1] In addition requires assessment of the use of 13 assistive devices or help from others with 8 activities examined content validity of the entire PROMIS item bank as well

[2] Score conversion tables available

[3] Assessment center pricing is available through http://www.healthmeasures.net/resource-center/about-us/pricing-for-services

Table 4.

Results from 3-Round Modified Delphi for Functional Status Assessment Measures

	HAQ				PROMIS		
	HAQ-DI	**MHAQ**	**MD-HAQ**	**HAQ-II**	**PF 10a**	**PF 20a**	**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	I	I	R **	R	R	I	I

R= Recommended, I=Inconclusive, N/A= not applicable (as measure included based on previous rounds of voting)

* Ratings reported by the number of participants voting in each range 1–3/4–6/7–9. The 1–9 Likert scale correspond to the following 1–3 “not recommended”; 2–4 “sometimes recommended”; 7–9 “essential to have”

** During review by the ACR Quality Measures Subcommittee the additional final recommendation of MD-HAQ for preferred use was based upon Delphi rating, feasibility, current use, and strength of its inclusion in the prior and concurrent ACR rheumatoid arthritis (RA) disease activity measure recommendations within the RAPID3.