Seminars in Arthritis and Rheumatism 55 (2022) 152029

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Validation studies of rheumatoid arthritis patient-reported outcome measures in populations at risk for inequity: A systematic review and analysis using the OMERACT summary of measurement properties equity table

Cheryl Barnabe [a,*], Aimée Wattiaux [b], Jennifer Petkovic [c], Dorcas Beaton [d], Beverley Shea [e], Regina Greer-Smith [f], Jenny Humphreys [g], Christie Bartels [b], Peter Tugwell [e], Valerie Umaefulam [a]

[a] Cumming School of Medicine, University of Calgary, 3330 Hospital Dr NW, Calgary AB T2N4N1, Canada
[b] University of Wisconsin School of Medicine and Public Health, 750 Highland Ave, Madison, WI 53726, USA
[c] Campbell and Cochrane Equity Methods Group, University of Ottawa, Canada
[d] Institute for Work & Health, Department of Occupational Therapy, University of Toronto, 400 University Ave, Suite 1800, Toronto ON, M5G 1S5, Canada
[e] Ottawa Hospital Research Institute, Department of Medicine and School of Epidemiology and Public Health, University of Ottawa, Canada
[f] Healthcare Research Associates, LLC/S.T.A.R. Initiative, USA
[g] Centre for Epidemiology Versus Arthritis, Division of Musculoskeletal & Dermatological Sciences, The University of Manchester, Oxford Rd, Manchester , M139PL, United Kingdom

ARTICLE INFO

Keywords:
Patient-reported outcome measures
Equity
Feasibility
Construct validity
Discriminant validity
OMERACT

ABSTRACT

Background: Existing patient-reported outcome measures (PROMs) in rheumatoid arthritis (RA) may be limited in their applicability to populations that are at risk for inequities. We conducted a systematic review to identify and rate evidence in the validation studies for PROMs in populations at risk for inequity.
Methods: A systematic review of MEDLINE and EMBASE was completed. The search strategy was developed to identify measurement property studies for PROMs of interest (selected pain, disease activity, global evaluation and quality of life scales) in patients with RA. We identified experimental, observational, and qualitative studies reporting analysis of feasibility, construct validity and discriminant ability metrics for populations at risk for inequity by various factors including race, ethnicity, culture or language; employment status; sex and gender identity; education level; socioeconomic status; social support; age; health literacy and disability. These were rated based on the OMERACT Summary of Measurement Properties Equity table.
Results: From 19,786 titles and abstracts screened, we identified 14 unique studies reporting validation metrics for pain ($n = 3$), DAS28-ESR or DAS28-CRP ($n = 2$), ACR20 ($n = 1$), patient global assessment ($n = 2$), EQ5D ($n = 4$), and PROMIS® ($n = 3$) by race ($n = 10$ studies), age ($n = 6$ studies), sex ($n = 5$ studies), education level ($n = 2$ studies), and disability, literacy, employment status, social support level and socioeconomic status ($n = 1$ study each). Five studies reported on feasibility, 12 reported construct validity metrics, and 4 studies reported on discriminant validity metrics. All studies by culture or language were rated as having good measurement property metrics. There was limited assessment of measurement property metrics for other populations at risk for inequity.
Conclusion: Our study highlights important gaps in patient representation in rheumatology research for accepted outcome measures. New outcome measures being developed for research purposes and clinical practice should ensure and report representation of patients from populations at risk for inequities in the testing of metrics of feasibility, construct validity and discriminant ability metrics.

* Corresponding author.
E-mail address: ccbarnab@ucalgary.ca (C. Barnabe).

https://doi.org/10.1016/j.semarthrit.2022.152029

Available online 19 May 2022

Introduction

Patient-reported outcome measures (PROMs) are a cornerstone in assessing the burden and impact of rheumatic diseases such as Rheumatoid Arthritis (RA) in daily clinical practice, clinical trials and longitudinal registries. It is uncertain how frequently and to what extent equity has been considered in the development, selection or testing of PROMs. In previous work, the Equity Special Interest Group of OMERACT (Outcome Measures in Rheumatology) has demonstrated how literacy and culture impact validated tools of the Health Assessment Questionnaire (HAQ) and Osteoarthritis Knee and Hip Quality of Life Questionnaires [1], and the differential effects of RA based on sex [2]. This collaboration also set a research agenda for including an equity lens in PROM development, selection and validation testing [2].

In order to include the many and intersecting [3,4] populations that are at risk for inequities, OMERACT's Equity Special Interest Group chose to use the PROGRESS-Plus framework [5] to identify characteristics that stratify health opportunities and outcomes. These are conceptualized as **P**lace of residence, **R**ace/ethnicity/culture/language, **O**ccupation, **G**ender/sex, **R**eligion, **E**ducation, **S**ocioeconomic status, and **S**ocial capital, with "Plus" referring to additional characteristics such as age, disability, and relationships where one person may have more power than another. We used this framework in developing this review, which sought to systematically identify published PROM validation studies conducted in populations with RA at risk for inequities, rate the available evidence using the OMERACT Summary of Measurement Properties Equity table [6], and highlight gaps in the available evidence.

Methods

Adherence to guidelines

The design and methods used for this review are reported in line with Preferred Reporting Items for Systematic Reviews and Meta-Analyses for reporting equity-focused systematic reviews (PRISMA-E) [7]. Our study was registered in Prospero on November 18, 2019 [8].

Search strategies and databases

A search strategy combining PROM measurement property studies [9], RA terms, and pre-selected PROM instruments was developed for MEDLINE Ovid (1946 to 21 November 2019) and EMBASE Ovid (1974 to 21 November 2019). The terms used in the MEDLINE search strategy are provided in Appendix 1. We applied a restriction to remove study designs that were not of interest, and did not apply restrictions on language of publication. Identified studies were imported into Covidence (Melbourne, Australia) for record management, and duplicate entries were removed.

Eligibility criteria and study selection

Titles and abstracts of studies retrieved were screened independently for meeting the eligibility criteria by a minimum of two reviewers (VU, AW, JH). Another reviewer (JP) resolved disagreements where needed. We identified studies with experimental, observational, or qualitative designs that reported on the validation of PROMs specifically in patients with RA in at least one of the population groups characterized in the PROGRESS-Plus acronym. Validation studies could address any of feasibility, construct validity, or discriminant ability (e.g., test/retest reliability, responsiveness, thresholds of meaning) metrics. The pre-selected PROM instruments of interest were pain (100 mm VAS or 5-point scale), physical function (HAQ and grip strength), disease activity measures (DAS28-ESR or DAS28-CRP, ACR20, Patient Global Assessment 100 mm VAS), quality of life (EQ5D), and any item banks of the PROMIS® instrument. Potentially eligible studies were retrieved for full text review, completed independently by two reviewers (VU, AW) with a third reviewer (JP) resolving disagreements.

Data extraction

Studies meeting the inclusion criteria were characterized for the publication year, study design, country of study, and characteristics of the study participants. We extracted information on the populations and the PROM instruments assessed in the study. We then extracted validation metrics into a developed OMERACT Summary of Measurement Properties (SOMP)-Equity table [6] which was adapted from the original OMERACT SOMP Table [9]. Two reviewers (VU and CB) extracted data from the included studies into standardized tables for each PROM. To assess feasibility, we extracted clear descriptions of patient or provider perspectives on the usability of the instrument, specific to burden, time, effort, translations, and cost of using the instrument. To assess construct validity, we extracted descriptions of the degree to which scores of the instrument relate to other measures, either clinical indicators or patient-reported items on other scales. For discriminant ability, we extracted descriptions of any metrics including test-retest reliability, responsiveness to change in state, and generation of meaningful thresholds of change in state.

Data analysis

A narrative synthesis was completed to summarize which validation metrics are available in various PROGRESS-Plus populations across the different PROMs, as well as the results of these validation studies. The available evidence was then assigned a rating according to the OMERACT 'traffic light scoring' method as described in their handbook [9] and which summarizes the certainty of effect for that metric. 'Green' reflects that the primary study concluded that the instrument met feasibility, construct validity or discriminant ability requirements for the construct tested. Items in 'Amber' are those where the measurement properties were potentially valid or had discriminant ability however with remaining uncertainty. Items in 'Red' did not have construct validity or discriminant ability in the population it was tested in. Items without evidence remain 'White' in the SOMP-Equity table. Each included study underwent a quality assessment using the COSMIN—OMERACT Good Methods Checklists, providing an overall rating for each individual construct (excluding feasibility) as either likely low risk of bias, with caution but may be used as evidence, or that the evidence should not be used [9]. If there was concern for bias and the evidence should not be used, it was also rated as 'Red'.

Results

Description of studies

The PRISMA flow diagram (Fig. 1) provides an overview of the study selection process. A total of 24,889 studies were imported into Covidence for screening and 5103 duplicate studies were removed. A total of 19,786 titles and abstracts were screened and 183 were retained for full text review. Overall, there were 48 relevant studies that assessed the selected measures across population groups. HAQ results have been accepted for publication, thus we present results for the other PROMs of interest [6].

We identified 14 unique studies reporting validation metrics on the PROMs of interest. This included 3 studies of the pain VAS scale [10–12], 2 studies on the DAS28-ESR and DAS28-CRP [13,14], 1 study on the ACR20 [15], 2 studies on the patient global assessment [12,16], 4 studies on the EQ5D [17–20], and 3 studies on the PROMIS instrument [21–23] (note: one study reported results for both the pain VAS scale and the patient global assessment [12]). No study reported on grip strength, nor were there any pain 5-point scale instrument studies. Twelve studies focused on construct validity ($n = 12$), with 5 studies

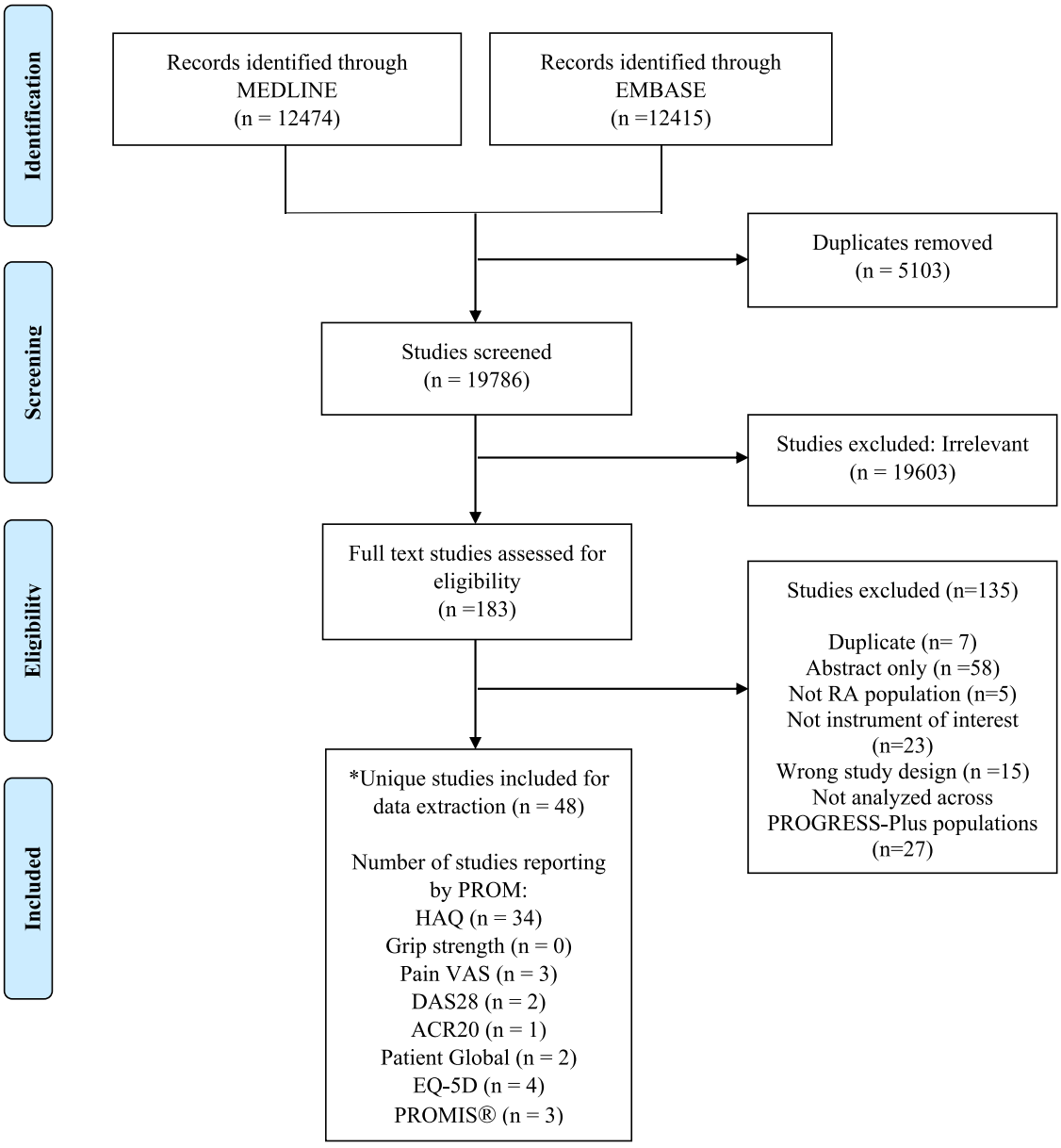

Fig. 1. PRISMA Study Flow Diagram.

reporting on feasibility and 5 on discriminant ability metrics. In the identified studies, some reported results for a single population at risk of inequity, whereas others reported results for multiple populations. In total, 10 studies reported language or cultural adaptations of existing instruments or were analysed for validity across various ethnicities. Six studies examined validity by age, 5 studies reported on validity in males and females or men and women, and 2 studies analysed validity by educational level. Health literacy, disability, socioeconomic status, social support level and employment status were each reported in 1 study. A summary of study characteristics and quality assessment is provided in Table 1.

A summary of findings is presented in the SOMP-Equity table (Table 2), with further detailed results summarized in the text.

Pain VAS

Escalante reported almost perfect agreement between English and Spanish versions (intraclass correlation coefficient=0.92), and the Spanish version completed prior to and then following the physician visit (within 60–180 min) was again almost perfect (intraclass correlation coefficient 0.95) [10]. Tamiya reported that the Japanese version was feasible to complete (readily understandable and easily filled out). Construct validity in this study was demonstrated by a correlation between higher levels of pain being associated with higher CRP levels (one way ANOVA $F = 6.09$, df=4, $p = 0.015$). Same-day (Pearson's correlation coefficient 0.85, 95%CI 0.78–0.88) and 7 day test-retest reliability (Pearson's correlation coefficient 0.64, 95%CI 0.43–0.78) were significantly correlated [11]. In Colombian patients, the suite of PROMs including the pain VAS were assessed to be easily completed for the majority of participants. By the spearman's rank correlation coefficient, the pain VAS was not well correlated to swollen joint count (0.323), DAS28-ESR (0.434), nor MD Global (0.314), and was only moderately correlated to tender joint count (0.508), CDAI (0.632) and SDAI (0.606) [12].

DAS28

Leeb and colleagues contrasted DAS28-ESR scores between females

Table 1
Patient-Reported Outcome Measure Validation Studies in Populations at Risk for Inequities.

PROM Instrument (s) Validated	First Author and Year	Country of study	Study Design	PROGRESS-Plus population	Study Participants	Key Participant Characteristics	Measurement properties studied	Quality Assessment
Pain VAS	Escalante 1996 [10]	USA	Cross-sectional	Spanish language	20 bilingual participants for equivalence study, and 23 English and 20 Spanish speaking monolingual patients for test-retest reliability	Subset characteristics not provided	Construct ValidityTest-Retest Reliability	Low risk of biasLow risk of bias
	Tamiya 2002 [11]	Japan	Cohort	Japanese culture/ language	145 female participants, with 47 participating in the reliability assessment	Mean age 53 years and disease duration 11 years	FeasibilityConstruct ValidityTest-Retest Reliability	n/aLow risk of biasLow risk of bias
	Amaya-Amaya 2012 [12]	Colombia	Cross Sectional	Colombian patients	135 participants receiving focus group instruction on PROMs and then relationship between disease activity and PROMs assessed by physicians	Mean age 54 years, median disease duration 12 years, 79% were female, 44% had low educational level, and 39% had low socioeconomic status	FeasibilityConstruct Validity	n/aLow risk of bias
DAS28	Leeb 2007 [13]	Austria	Cross sectional	Sex, Age	557 participants to determine influence of sex, age and disease duration on DAS28-ESR scores	Median age 64 years, median disease duration 48 months, 78% female. Male patients had a significantly shorter disease duration	Construct Validity	Low risk of bias
	Park 2012 [14]	Korea	Cohort	Korean patients	223 participants to determine thresholds for disease activity states between DAS28-ESR and DAS28-CRP	Mean age 53 years, 89% were female. Mean disease duration 13 years in females	Construct ValidityThresholds of Meaning	Low risk of biasLow risk of bias
ACR20	Ward 2014 [15]	USA	Cohort	Age, sex, White, Black and Hispanic ethnicity	250 participants with active RA undergoing treatment initiations or escalations	Mean age 51 years, mean disease duration 10 years, 78% female, 41% White non-Hispanic, 29% Hispanic, 22% Black, 7% Asian and 1% multiethnicity	Responsiveness	Low risk of bias
Patient Global Assessment	Hirsh 2010 [16]	USA	Cross sectional	Health literacy	110 participants recruited from an urban "safety net" system with a high proportion of minority ethnicities and persons without health insurance but with English as their primary language	Mean age 53 years, mean disease duration 13 years, 79% female, mean 12 years of education, 27% 'Caucasian', 16% Black, 11% American Indian, 3% Asian, 19% currently employed, 65% disabled	FeasibilityConstruct Validity	n/aLow risk of bias
	Amaya-Amaya 2012 [12]	Colombia	Cross Sectional	Colombian patients	135 participants receiving focus group instruction on PROMs and then relationship between disease activity and PROMs assessed by physicians	Mean age 54 years, median disease duration 12 years, 79% were female, 44% had low educational level, and 39% had low socioeconomic status	FeasibilityConstruct Validity	n/aLow risk of bias
EQ5D	Hurst 1997 [17]	United Kingdom	Cohort (results extracted are cross-sectional)	Disability, socioeconomic status, social support, employment, education, age	233 participants stratified by functional class; persons from class 4 were recruited from hospitals, primary care and nursing homes	81% female with mean age 55 and disease duration 14 years; 19% male with mean age 58 years and disease duration of 9 years. Age and disease duration increased with worsening functional class. Owner-occupied property served as a proxy for socioeconomic status (65%), living with a spouse or partner (67%) or living with a 'carer' were indicators of social support. 71% were unemployed	Construct ValidityThresholds of Meaning	Low risk of biasCaution but evidence may be used
	Salaffi 2011 [18]	Italy	Cross sectional	Age, Education	583 participants	Mean age 58 years, 75% female, 74% with low educational level	Construct Validity	Low risk of bias
	Ferreira 2013[19]	Portugal	Cross sectional	Portuguese language/ culture	104 participants from a registry	Characteristics not reported	Construct Validity	Should not be used as evidence
	Munchey 2018[20]	Thailand	Cross sectional	Thai language, Age, Sex	221 participants	Median age 58 years, median disease duration 42 months, 78% female, 57% were aged 41 to 60 years, 67% primary education or lower, 81% with health coverage, 33% not working and 33% working in agriculture, 83% low income	Construct Validity	Low risk of bias
PROMIS®	Oude Voshaar 2012[21]	Netherlands	Cohort	Dutch language/culture	15 participants (5 additional participants had osteoarthritis or psoriatic arthritis)	Subset characteristics not provided	Feasibility	n/a

(*continued on next page*)

Table 1 (*continued*)

PROM Instrument (s) Validated	First Author and Year	Country of study	Study Design	PROGRESS-Plus population	Study Participants	Key Participant Characteristics	Measurement properties studied	Quality Assessment
	Mahmoud 2019[22]	Egypt	Cross sectional	Arabic language/ culture, Sex	100 participants	Mean age 42 years, mean disease duration 6 years, 94% females, 50% with basic education, 31% with secondary education and 19% high education	Construct validity	Low risk of bias
	Crins 2020 [23]	Netherlands	Cohort	Dutch-Flemish language, Age, Sex	2029 participants for the PROMIS®-Pain Interference item bank and 1554 participants for the PROMIS®-Pain behavior item bank	PROMIS®-Pain Interference item bank participants: Mean age 59 years, 69% female, 72% married or living common law, 33% with college or advanced degree, 38% with some college, 17% with high school and 12% with less than high school, 42% retired or unemployed, 30% with social benefitsPROMIS®-Pain Behaviour item bank participants: Mean age 59 years, 69% female, 74% married or living common law, 32% with college or advanced degree, 38% with some college, 17% with high school and 13% with less than high school, 38% retired of unemployed, 29% with social benefits	Construct validity	Low risk of bias

Legend: PROMs patient-reported outcomes measures; VAS visual analogue scale; DAS28 disease activity score based on 28 joint counts; ESR erythrocyte sedimentation rate; CRP C-reactive protein. All values in 'Key Participant Characteristics' rounded to closest integer value, n/a not applicable.

and males. They determined that there were higher mean DAS28-ESR scores in females relative to males (3.66 (SD 0.57) vs 3.01 (SD 1.12), $p<0.0001$) driven by higher swollen and tender joint counts, physician and patient global evaluation scores, and ESR levels. However, as CRP levels and pain levels were similar between females and males, the authors suggest that the DAS28-ESR may be affected by sex rather than underlying disease activity, and raised the question if thresholds of disease activity should be reconsidered although this was not formally studied. This study also examined the influence of age on DAS28-ESR scores, with a borderline significant relationship (linear regression $r = 0.130$, $p = 0.02$) [13]. In a study of Korean patients, DAS28-ESR was highly correlated with the DAS28-CRP (by linear regression estimate 0.93, $p<0.0001$), and using receiver operating characteristic curves it was determined that corresponding DAS28-CRP cutpoints should be 2.19 for remission (vs 2.6 for the DAS28-ESR), 2.60 for low disease activity (vs 3.2 for DAS28-ESR) and 4.07 for high disease activity (vs 5.1 for DAS28-ESR), with high sensitivities (range 0.89–0.91), specificities (range 0.82–0.96) and accuracy (range 0.88–0.95) [14].

ACR20

In an analysis of ACR20 improvement with treatment, specificities and positive predictive values were similar in patient subgroups defined by age, sex and ethnicity. However, although not statistically significant the mean sensitivities for response were lower for patients ≥60 years (0.42, 95%CI 0.27–0.59 vs 0.61, 95%CI 0.52–0.70 for <60 years), male patients (0.45, 95%CI 0.29–0.64 vs 0.60, 95%CI 0.51–0.69 for females), and for black (0.51 95%CI 0.34–0.69) and white (0.51, 95% CI 0.37–0.64) patients relative to Hispanic (0.64, 95%CI 0.50–0.77) patients [15].

Patient global assessment

A study by Hirsh assessed the variations of patient global assessments as used in the MDHAQ and the DAS28 tools, and how health literacy impacted the discrepancy between patient-reported values to the provider global assessment. The Gunning-Fox Index estimates of literacy for the global assessment scales were 14.06 and 11.47 respectively. There was no significant within-patient difference between the patient global assessment scores, and provider global assessment scores were consistently lower. The difference between patient and provider global assessment scores narrowed with higher health literacy levels as measured by 'Rapid Estimate of Adult Literacy in Medicine' (REALM) and 'Short Test of Functional Health Literacy in Adults' (S-TOFHLA) instruments in linear regression (all estimates statistically significant), even after controlling for biological disease-modifying drug use, years of education, sex and age in the analysis (S-TOFHLA coefficient −0.39706, 95%CI −0.75795 to −0.03618, $p = 0.031$; REALM coefficient not provided) [16]. Amaya-Amaya's study of Colombian patients found the strongest degrees of correlation between the patient global assessment and the disease activity indices CDAI and SDAI (spearman's rank correlation coefficients 0.754 and 0.725 respectively), moderate correlation with the DAS28 score and tender joint counts (spearman's rank correlation coefficients 0.517 and 0.583 respectively) with low correlation to the swollen joint count (spearman's rank correlation coefficient 0.396) and provider global assessment score (agreement by weighted kappa=0.026) [12].

EQ5D

The analysis by Hurst demonstrated that median unweighted EQ5D ratings ($EQ5D_{profile}$) across the five domains worsened progressively with worsening functional class (increasing disability) (Kruskal-Wallis test $p<0.001$), and the spearman rank correlation coefficient between functional class and $EQ5D_{utility}$ was −0.74 and with $EQ5D_{vas}$ −0.55. $EQ5D_{utility}$ was better with social support ($p<0.05$ for living with

Table 2
OMERACT Summary of Measurement Properties (SOMP)-Equity Extension for Patient-Reported Outcome Measures.

	Feasibility	Truth Construct Validity	Discriminant Ability Test-Retest Reliability	Responsiveness	Thresholds of Meaning
Place of Residence					
Race, Ethnicity, Culture, Language	Pain VAS (Japanese[11])PGA (Colombian[12])PROMIS® Physical Function Item Bank (Dutch[21])	Pain VAS (Spanish[10], Japanese[11])DAS28-CRP (Korean[14])PGA (Colombian[12])$EQ5D_{utility}$ (Thai[20])PROMIS® Physical Function Item Bank (Arabic[22])PROMIS® Pain Interference Item Bank (Arabic[22])PROMIS® Pain Intensity Item Bank (Arabic[22]) PROMIS® Fatigue Item Bank (Arabic[22])PROMIS® Sleep Item Bank (Arabic[22])PROMIS® Anxiety Item Bank (Arabic[22])PROMIS® Depression Item Bank (Arabic[22])PROMIS® Pain Interference Item Bank (Dutch-Flemish[23])PROMIS® Pain Behaviour Item Bank (Dutch-Flemish[23])	Pain VAS (Spanish[10], Japanese[11])		
	Pain VAS (Colombian[12])			ACR20 (White, Black, Hispanic [15])	DAS28-CRP (Korean[14])
		Pain VAS (Colombian[12])$EQ5D_{utility}$ (Portuguese[19])			
Occupation		$EQ5D_{utility}$[17]$EQ5D_{vas}$[17]			
Gender/Sex Identity		PROMIS® Pain Behaviour Item Bank (Dutch-Flemish [23])			
				ACR20[15]	
		DAS28-ESR[13]$EQ5D_{utility}$ (Thai[20])			
Religion					
Education Level		$EQ5D_{utility}$[17]$EQ5D_{vas}$[17] $EQ5D_{profile}$[18]			
Socioeconomic status		$EQ5D_{utility}$[17] $EQ5D_{vas}$[17]			
Social support level		$EQ5D_{utility}$[17] $EQ5D_{vas}$[17]			
Age		$EQ5D_{profile}$[18]$EQ5D_{utility}$[17]$EQ5D_{vas}$[17] DAS28-ESR[13] $EQ5D_{utility}$ (Thai[20])PROMIS® Pain Behaviour Item Bank (Dutch-Flemish[23])	ACR20[15]		
Health literacy	PGA[16]	PGA[16]			
Disability		$EQ5D_{profile}$ [17]			$EQ5D_{utility}$ [17] $EQ5D_{vas}$ [17]

Legend: VAS visual analogue scale; PGA patient global assessment; DAS28 disease activity score based on 28 joint counts; CRP C-reactive protein. Items in green are those where the primary studies concluded that the instrument met feasibility, construct validity or discriminant ability requirements for the construct tested and good evidence supported this property, passing the element of the filter. Items in amber are those where the measurement properties were potentially valid or had discriminant ability however with remaining uncertainty due to the limited evidence. Items in red indicate did not have construct validity or discriminant ability in the population it was tested in or poor quality evidence. Items without evidence are in white.

spouse/partner and $p<0.01$ for living with a 'carer') whereas $EQ5D_{vas}$ was better for persons living without social supports. $EQ5D_{utility}$ was significantly higher for persons with higher socioeconomic status, and $EQ5D_{vas}$ was not different by socioeconomic status. Employment status was protective, with both $EQ5D_{utility}$ and $EQ5D_{vas}$ significantly higher than unemployed persons ($p<0.001$). Significant correlations were also seen with years of education ($EQ5D_{utility}$ $R = 0.33$; $EQ5D_{vas}$ $R = 0.28$) and age ($EQ5D_{utility}$ $R= -0.29$; $EQ5D_{vas}$ $R= -0.17$). The $EQ5D_{utility}$ value discriminated well between each functional class whereas the $EQ5D_{vas}$ did not discriminate well between worse functional classes (3 and 4) [17]. In Italian patients (but with weighting of $EQ5D_{profile}$ scores to a United Kingdom population) there was a significant correlation of the score with age (spearman rank correlation coefficient -0.089, $p = 0.031$) but not educational level (coefficient 0.026, $p = 0.536$) [18]. The Portuguese version of the EQ5D was tested in 104 patients with RA; the manuscript states there were strong and inverse correlations between $EQ5D_{utility}$ measures and symptom-associated dimensions, affect and social interaction of the AIMS2-SF but results were not presented [19]. In a group of RA patients completing the Thai version of the EQ5D, there was a strong correlation between $EQ5D_{utility}$ and the HAQ score (spearman's rank correlation coefficient -0.65), and a moderate correlation between $EQ5D_{vas}$ and the HAQ score (coefficient -0.39), both $p<0.001$). Internal consistency was determined to be acceptable (Cronbach alpha coefficient 0.75). There were no significant differences in $EQ5D_{utility}$ or $EQ5D_{vas}$ scores by age or sex [20].

PROMIS

Oude Voshaar and colleagues completed a Dutch translation and cross-cultural validation of the 124 items of the PROMIS tool's physical function item bank. After an intensive process of forward translation, reconciliation, back translation and then with input of an expert committee, a pre-final version of the tool was administered to patients predominantly with RA. Feasibility was reported qualitatively, with the authors reporting that questions were well understood by patients, supported by think aloud observation. They noted a constant speed in answering the items, with the exception of a single question (item 31, '*Are you able to lift one pound (a full pint container) to shoulder level without bending your elbow?*) that was subsequently further revised for clarity [21]. Arabic translation and cross-cultural adaptation of several PROMIS items for physical function, pain intensity, pain interference, fatigue, sleep, anxiety and depression) was completed followed by examining the correlations between the PROMIS domains with the HAQ, patient global assessment, and disease activity measures (DAS28-ESR and CDAI). All PROMIS domains studied significantly correlated with these other disease measures [22]. The Dutch-Flemish versions of the PROMIS Pain Interference and Pain Behaviour item banks were administered to 2029 and 1554 patients respectively. Despite similar levels of pain behaviour, Flemish patients were slightly more likely to endorse item 16 "*When I was in pain I appeared upset or sad*" than Dutch patients. A single item of the Pain Behaviour item bank was differential

by sex, with women with similar levels of pain behaviour as men being slightly more likely to endorse item 27 *"I had pain so bad it made me cry."* However, this had negligible impact on the PROMIS T-scores. Item parameters were equivalent in patients differing in age. The PROMIS Pain Interference correlated strongly with the Dutch-Flemish PROMIS Global Health Pain intensity (Pearson correlation coefficient $r = 0.80$), SF36-Physical Function 10 (Pearson correlation coefficient $r = -0.71$) and HAQ-DI (Pearson correlation coefficient $r = 0.71$) instruments. The PROMIS Pain Behaviour item bank correlated strongly with the Global Health Pain intensity (Pearson correlation coefficient $r = 0.61$) [23].

Discussion

The intent of this systematic review was to extend our work on which PROMs used in RA assessment had been validated in different populations at risk for inequities. The experience of RA may be highly different between population groups, and the inclusion of diverse representation as well as focused testing in particular populations will ensure the final instrument measures the intended concept despite variations in individual characteristics and experiences [24]. Despite a broad systematic search strategy, we identified relatively few studies. This is of great concern, as validity assessment should be a critical procedure in the development and testing of all outcome measures, especially those related to patient status. We discuss our findings to support these statements here.

Lack of representativeness in populations included in prom research

While language and cross-cultural validation studies generally resulted in good feasibility, construct validity and discriminant ability metrics, we only identified a small number of studies conducted with limited numbers of countries represented and few patients enroled. One prior report had identified 39 studies that described translation, cultural adaptation and/or cross-cultural validity of the HAQ and its derivatives and the PROMIS functional status assessment measures in RA supporting interest in language translations and cultural adaptations that are being conducted in rheumatology research [25]. However, there is less assessment of the validity of PROMs by other population characteristics such as age, education level, employment, socioeconomic status, disability, health literacy or social supports. Notably, no studies examined the impacts of diversity in gender identity and expression nor sexual orientation on PROMs. There are few descriptions of disease and care experience in the RA or broader rheumatology literature for sexual and gender minorities. Yet, from an American study that included persons with a self-reported diagnosis of all types of arthritis, lesbian and bisexual women had 55% and 54% higher odds of disability respectively than heterosexual women [26]. There are significant implications of this lack of validation of PROMs in diverse populations. As PROMs are used to support access to advanced treatments, there is a real risk that patients from these populations may be systematically excluded from accessing necessary therapy, and thus increasing outcome inequities. Altogether, this supports that the rheumatology community must begin to evaluate the relevance and property measurement characteristics of PROMs for existing and new instruments, to ensure they accurately reflect the experiences of marginalized populations at risk for inequities within society, healthcare, and specifically in arthritis care.

Hierarchy of data collection approaches and a paucity of longitudinal data collection to determine responsiveness

Another finding of this review is the focus of the literature on quantitative aspects of construct validity, with few studies characterizing feasibility and discriminant ability metrics of reliability, responsiveness and thresholds of meaningful change longitudinally. We were unable to identify any significant volume of qualitative research associated with the development of PROMs, contrary to the guidance that patient-centred PROMs must be meaningful and understandable to diverse population groups when administered, which requires direct patient input and perspective during development, revisions, and validation [27]. Qualitative data, particularly via cognitive interviewing, are vital for establishing the content validity of the PROMs instrument, ensuring the instrument adequately covers the domain of interest, and is a good research practice for measure development [24,28]. Secondly, while there are recognized challenges with engaging patients from populations that face inequities in research, responsiveness estimates are important to detect changes in health over time that matter to the patient [29]. The rheumatology research community involved in property measurement studies should consider approaches that uphold community engagement principles to obtain robust metrics.

Strengths and limitations

We highlight gaps in the existing rheumatology research literature, aligning with a broader societal focus on inequities that exist in the population on the basis of race and ethnicity, access to resources, and sexual orientation as a few examples. We urge caution in the assessments we present in the SOMP-Equity Table, as for many instruments there is a single study reported upon which we drew conclusions. Prior development and validation studies of PROMs may in fact have had population diversity, but our work did not formally assess the population characteristics enroled in those studies if not formally analysed. This could be done with post-hoc analyses of prior studies, and future measures should be explicit in describing the diversity of enrollees and/or conducting population-specific validation studies.

Conclusion

Our review identified important gaps in representation of populations at risk for inequities in rheumatology research. We encourage researchers to carefully consider their approach to PROM development, specifically by including patients in priority populations for diversity and equity, and consider appropriate instrument selection and development methods.

Funding

This work was supported by the University of Calgary [Umaefulam, Eyes High Postdoctoral Scholarship], and the Canadian Institutes of Health Research [Barnabe, Foundation Scheme Grant and Canada Research Chair in Rheumatoid Arthritis and Autoimmune Diseases]

Declaration of Competing Interest

None.

Appendix 1: Search Strategy in MEDLINE

#1 POPULATION:“Rheumatoid arthritis” OR “RA” OR “inflammatory arthritis” OR “inflammatory polyarthritis” OR “early inflammatory arthritis” OR “early arthritis”#2 INSTRUMENT:“Pain VAS” OR “Pain Visual Analogue Scale” OR “Pain scale” OR “Health Assessment Questionnaire” OR “HAQ” OR “Grip strength” OR “Patient Global Assessment” OR “Patient Global” OR “DAS28” OR “DAS” OR “Disease activity score” OR “Disease activity” OR “ACR20” OR “EQ5D” “Quality of life” OR “QoL”(numeric*[tiab] AND rating [tiab] AND (scale[tiab] OR score[tiab])) OR numeric scale[tiab] OR nrs[tw] OR nprs[tw]#3 MEASUREMENT PROPERTIES:(instrumentation[sh] OR methods[sh] OR "Validation Studies"[pt] OR "Comparative Study"[pt] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment(health care)"[MeSH] OR "outcome assessment"[tiab] OR "outcome measure*"[tw] OR "observer variation"[MeSH]OR "observer variation"[tiab] OR "Health Status Indicators"[Mesh] OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR "coefficient of variation"[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internalconsistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item [tiab] AND (correlation*[tiab]OR selection*[tiab] OR reduction*[tiab])) OR agreement[tw] OR precision[tw] OR imprecision[tw] OR "precisevalues"[tw] OR test-retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab]))OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater [tiab] OR intertester[tiab]OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer [tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant [tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab]OR kappa’s[tiab] OR kappas[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] OR result[tw] OR results[tw] OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR "factor analysis"[tiab] OR "factor analyses"[tiab] OR "factor structure"[tiab] OR "factorstructures"[tiab] OR dimension* [tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab]OR analyses[tiab])) OR "item discriminant"[tiab] OR "interscale correlation*"[tiab] OR error[tiab] OR errors[tiab]OR "individual variability"[tiab] OR "interval variability"[tiab] OR "rate variability"[tiab]OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR "minimal detectable concentration"[tiab] OR interpretab*[tiab] OR((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important [tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab]OR difference[tiab])) OR (small*[tiab] AND (real [tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] ORIRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab]OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab])#4("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[PublicationType] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legalcases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[PublicationType] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popularworks"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[PublicationType] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type])NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])#5#1 AND #2 AND #3#6#5 NOT #4

References

[1] Petkovic J, Epstein J, Buchbinder R, et al. Toward ensuring health equity: readability and cultural equivalence of OMERACT patient-reported outcome measures. J Rheumatol 2015;42:2448–59.

[2] Petkovic J, Barton JL, Flurey C, et al. Health equity considerations for developing and reporting patient-reported outcomes in clinical trials: a report from the OMERACT equity special interest group. J Rheumatol 2017;44:1727–33.

[3] Crenshaw K. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum; 1989. p. 139–68.

[4] Nixon SA. The coin model of privilege and critical allyship: implications for health. BMC Public Health 2019;19:1637.

[5] O’Neill J, Tabish H, Welch V, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol 2014;67:56–64.

[6] Petkovic J, Umaefulam V, Wattiaux A, et al. Development of an extension of the OMERACT Summary of Measurement Property table to capture equity considerations: sOMP-Equity. Semin Arthritis Rheum 2021.

[7] Welch V, Petticrew M, Tugwell P, et al. PRISMA-Equity 2012 extension: reporting guidelines for systematic reviews with a focus on health equity. PLoS Med 2012;9: e1001333.

[8] Humphreys J., Umaefulam V., Barnabe C., et al. Equity of outcome measures in rheumatoid arthritis. November 18 ed. PROSPERO2019:CRD42019157462.

[9] Boers M, Kirwan JR, Tugwell P, et al. The OMERACT Handbook. 2019.

[10] Escalante A, Galarza-Delgado D, Beardmore TD, et al. Cross-cultural adaptation of a brief outcome questionnaire for Spanish-speaking arthritis patients. Arthritis Rheum 1996;39:93–100.

[11] Tamiya N, Araki S, Ohi G, et al. Assessment of pain, depression, and anxiety by visual analogue scale in Japanese women with rheumatoid arthritis. Scand J Caring Sci 2002;16:137–41.

[12] Amaya-Amaya J, Botello-Corzo D, Calixto OJ, et al. Usefulness of patients-reported outcomes in rheumatoid arthritis focus group. Arthritis 2012;2012:935187.

[13] Leeb BF, Haindl PM, Maktari A, Nothnagl T, Rintelen B. Disease activity score-28 values differ considerably depending on patient’s pain perception and sex. J Rheumatol 2007;34:2382–7.

[14] Park SY, Lee H, Cho SK, Choi CB, Sung YK, Bae SC. Evaluation of disease activity indices in Korean patients with rheumatoid arthritis. Rheumatol Int 2012;32: 545–9.

[15] Ward MM, Guthrie LC, Alba MI. Brief report: rheumatoid arthritis response criteria and patient-reported improvement in arthritis activity: is an American College of Rheumatology twenty percent response meaningful to patients? Arthritis & rheumatology 2014;66:2339–43.

[16] Hirsh JM, Boyle DJ, Collier DH, Oxenfeld AJ, Caplan L. Health literacy predicts the discrepancy between patient and provider global assessments of rheumatoid arthritis activity at a public urban rheumatology clinic. J Rheumatol 2010;37: 961–6.

[17] Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol 1997;36:551–9.

[18] Salaffi F, Carotti M, Ciapetti A, Gasparini S, Grassi W. A comparison of utility measurement using EQ-5D and SF-6D preference-based generic instruments in patients with rheumatoid arthritis. Clin Exp Rheumatol 2011;29:661–71.

[19] Ferreira PL, Ferreira LN, Pereira LN. Contribution for the validation of the Portuguese version of EQ-5D. Acta Med Port 2013;26:664–75.

[20] Munchey R, Pongmesa T. Health-related quality of life and functional ability of patients with rheumatoid arthritis: a study from a tertiary care hospital in Thailand. Value Health Reg Issues 2018;15:76–81.

[21] Oude Voshaar MA, Ten Klooster PM, Taal E, Krishnan E, van de Laar MA. Dutch translation and cross-cultural adaptation of the PROMIS® physical function item bank and cognitive pre-test in Dutch arthritis patients. Arthritis Res Ther 2012;14: R47.

[22] Mahmoud GA, Rady HM, Mostafa AM. Cross cultural adaptation and validation of an Arabic version of selected PROMIS measures for use in rheumatoid arthritis patients. The Egyptian Rheumatologist 2019;41:177–82.

[23] Crins MHP, Terwee CB, Westhovens R, et al. First Validation of the Full PROMIS Pain Interference and Pain Behavior Item Banks in Patients With Rheumatoid Arthritis. Arthritis Care Res 2020;72:1550–9.
[24] Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2–assessing respondent understanding. Value Health 2011; 14:978–88.
[25] Kulhawy-Wibe SC, Zell J, Michaud K, et al. Systematic Review and Appraisal of the Cross-Cultural Validity of Functional Status Assessment Measures in Rheumatoid Arthritis. Arthritis Care Res 2020;72:798–805.
[26] Fredriksen-Goldsen KI, Kim HJ, Barkan SE. Disability among lesbian, gay, and bisexual adults: disparities in prevalence and risk. Am J Public Health 2012;102: e16–21.
[27] Basch E, Abernethy AP, Reeve BB. Assuring the patient centeredness of patient-reported outcomes: content validity in medical product development and comparative effectiveness research. Value Health 2011;14:965–6.
[28] US Department of Health and Human Services. Food and Drug Administration. 2009. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), (CDRH) CfDaRH. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labelling Claims. Vol 2021.
[29] Gaafary M. A guide to PROMs methodology and selection criteria. In: El Miedany Y, editor. Patient reported outcome measures in rheumatic diseases. Switzerland: Springer International Publishing Switzerland; 2016. p. 21–58.