Seminars in Arthritis and Rheumatism 51 (2021) 1300–1310

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

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

Development of an extension of the OMERACT Summary of Measurement Properties table to capture equity considerations: SOMP-Equity

Jennifer Petkovic[a,*], Valerie Umaefulam[b], Aimée Wattiaux[c], Christie Bartels[c], Cheryl Barnabe[d], Regina Greer-Smith[e], Catherine Hofstetter[f], Lara Maxwell[g], Beverley Shea[h], Jennifer Barton[i], Alex Young Soo Lee[g], Jennifer Humphreys[j], Dorcas Beaton[k,1], Peter Tugwell[l,1]

[a] Bruyère Research Institute, University of Ottawa, Ottawa, ON, Canada
[b] Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[c] University of Wisconsin School of Medicine and Public Health Madison, Wisconsin, USA
[d] Departments of Medicine and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[e] Healthcare Research Associates, LLC/The S.T.A.R. Initiative, Los Angeles, CA 90033, USA
[f] OMERACT Patient Research Partner, Toronto, Canada
[g] Faculty of Medicine, University of Ottawa, Ottawa, Canada
[h] Ottawa Hospital Research Institute, School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada
[i] Oregon Health and Science University, Portland, OR, USA
[j] Centre for Epidemiology Versus Arthritis, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, and NIHR Manchester Biomedical Research Centre
[k] Institute for Work and Health and Institute for Health Policy Management and Evaluation, University of Toronto, Toronto, Canada
[l] Department of Medicine, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Keywords:
OMERACT
Patient-reported outcome measures
equity
inequalities
social determinants of health

ABSTRACT

Objective: To develop an equity extension of the OMERACT Summary of Measurement Properties (SOMP) Table, SOMP Equity to describe whether a patient reported outcome measure (PROM) works well among patients of diverse languages and cultures, education levels, and other population characteristics.
Methods: We used the PROGRESS-Plus framework to categorize equity characteristics assessed in trials of PROM. PROGRESS refers to Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education, Socioeconomic status, and Social Capital, while the 'plus' captures additional characteristics, such as age. We pilot tested our SOMP Equity Extension using the Health Assessment Questionnaire (HAQ) as a prototypical PROM.
Results: The SOMP Equity Extension retains the same columns as the original OMERACT SOMP (domain match, feasibility, construct validity, test-retest reliability, longitudinal construct validity, clinical trial discrimination, thresholds of meaning) but uses the PROGRESS-Plus characteristics as rows. We found several examples of studies of the HAQ which had assessed one or more PROGRESS-Plus characteristics.
Conclusions: The most commonly reported equity considerations were related to language. OMERACT Equity virtual meeting participants were polled and they indicated that the SOMP Equity Extension is useful for highlighting and tracking equity considerations for OMERACT Core Outcome Measurement Instruments.

Background

Equity refers to the absence of unfair and avoidable differences in health outcomes [1]. The (Outcome Measures in Rheumatology) OMERACT-Equity Working Group uses the acronym PROGRESS-Plus to identify socially stratifying factors which may contribute to differences in opportunities for health. PROGRESS refers to: Place of residence, Race/ethnicity/culture/language, Occupation, Gender/sex, Religion, Education and literacy, Socioeconomic status, and Social capital [2]. The Plus includes additional characteristics which may

* Corresponding author.
E-mail addresses: jennifer.petkovic@uottawa.ca (J. Petkovic), valerie.umaefulam@ucalgary.ca (V. Umaefulam), wattiaux@medicine.wisc.edu (A. Wattiaux), cb4@medicine.wisc.edu (C. Bartels), ccbarnab@ucalgary.ca (C. Barnabe), healthcareresearch@sbcglobal.net (R. Greer-Smith), mcfence@on.aibn.com (C. Hofstetter), lmaxwell@uottawa.ca (L. Maxwell), bevshea35@gmail.com (B. Shea), bartoje@ohsu.edu (J. Barton), alexyslee1118@gmail.com (A.Y.S. Lee), jenny.humphreys@manchester.ac.uk (J. Humphreys), dorcas.beaton@gmail.com (D. Beaton), ptugwell@uottawa.ca (P. Tugwell).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.semarthrit.2021.09.005
0049-0172/

contribute to health inequities, such as age, disability, and power dynamics.

To varying degrees, these characteristics are especially important in the development and implementation of Patient Reported Outcome Measures (PROMs) as they may affect readability, comprehensibility, and cultural appropriateness of instruments, as previously demonstrated [3]. Failure to consider potential differences related to these characteristics may lead to measurement errors or reduced generalizability. This can affect our ability to accurately evaluate the effect of interventions across populations with rheumatic diseases, including disadvantaged and underrepresented groups, and may contribute to increasing inequities. Given the increased emphasis on diversity and inclusion being demanded by funding agencies for the spectrum of representative individuals entered in clinical trials, these equity aspects are especially important for clinical trials that use these PROM instruments as outcome measures.

Endorsement by OMERACT requires that each instrument must pass the OMERACT Filter of 'Truth', 'Discrimination' and 'Feasibility' [4,5]. The OMERACT Summary of Measurement Properties (SOMP) table (see Table 1) provides a visual summary assessment of these measurement criteria from each study as follows: Truth (domain match, construct validity), Discrimination (test-retest reliability, longitudinal construct validity, clinical trial discrimination, thresholds of meaning), and Feasibility. Individual studies are first assessed for risk of bias and those found to have low or some concerns are further assessed to determine whether they demonstrate adequate results for the measurement property.

The literature gathered for each measurement property is synthesized and the bottom cell of each column is assigned a rating of either GREEN (good evidence supporting this property, passes this element of the Filter), AMBER (some caution, or perhaps only one study on that property, but good enough to move forward) or RED (stop, evidence against this property or only poor-quality evidence). If there is no adequate quality evidence available on that property, it can be assigned a WHITE rating and await the creation of that evidence and future update of the rating (see Table 1) [6]. Outcome Instruments are awarded 'OMERACT Endorsement' for the overall result across all studies and their participants to receive a provisional (yellow) or final (green) rating.

The Original SOMP assesses overall results across all study participants but does not demonstrate whether a Domain Instrument works well for different settings, languages, cultures, education levels, and other population characteristics. Therefore, the OMERACT Equity Working Group elected to explore the idea of a SOMP Equity Extension tool to address issues of equity in assessing Core Set outcome instruments and demonstrate this with an example of a PROM to show whether it works well for different languages and cultures, different levels of education, and other aspects of diversity. The goal of the OMERACT SOMP-Equity extension table is to indicate that Core Outcome Measurement Set instruments have demonstrated that the OMERACT Filter criteria of Truth, Discrimination and Feasibility have also been met among patients from disadvantaged and/or underrepresented groups.

This work was discussed at the OMERACT Equity Special Interest Group session in November 2020 for which there were 47 attendees including 7 patients, two of whom are authors on this paper.

Methods

We assembled a Steering Group to inform the research process. This Steering Group included two equity working group co-chairs, two rheumatologist members of the working group, two patient research partners, three research fellows, two OMERACT senior methodologists, and the Chair of the OMERACT Handbook Group.

We decided to use the PROGRESS-Plus Framework for categorizing Equity characteristics with which members of this Working Group have experience.

As an exemplar, we used a prototypical PROM, the Health Assessment Questionnaire (HAQ-DI) in the initial development of the SOMP Equity Extension. The HAQ-DI is a widely-used patient-reported outcome measure developed for patients with rheumatic diseases to assess pain and disability [7]. It is included in the OMERACT Core Set for Rheumatoid Arthritis and it has been adapted and translated for use in many countries [8,9]. The HAQ includes questions related to whether a patient has been able to do the following activities over the past week:

- Upper limb

◯ Dress yourself, including tying shoelaces and doing buttons?
◯ Shampoo your hair?
◯ Cut your meat?
◯ Lift a full cup or glass to your mouth?
◯ Open a new milk carton?
◯ Wash and dry your entire body?
◯ Reach and get a 5 lb object from just above your head?
◯ Open car doors?

- Lower limb

◯ Stand up from an armless chair?
◯ Get in and out of bed?
◯ Walk outdoors on flat ground?
◯ Climb up 5 steps?
◯ Take a tub bath?
◯ Get on and off the toilet?
◯ Get in and out of car?
◯ Do chores such as vacuuming and yard work?
◯ Bend down and pick up clothing from the floor?

In a population for which these are not common activities (e.g. across cultures), patients cannot accurately assess their pain and disability if the questions do not have relevance for their daily experiences. For example [3]:

- 'Taking a tub bath' where tub baths are rare.
- 'Lifting a 5 lb object such as a bag of sugar', where sugar does not come in bags.
- 'Open a new milk carton', where milk does not come in cartons.
- 'Cutting meat', if patient is vegetarian.

Literature search: we searched MEDLINE and EMBASE databases with no date, time, or language restrictions to identify experimental, observational analytical, and qualitative studies on the development and assessment of the HAQ. Two members of the steering group screened the titles/abstracts of the references identified in our search and, independently, in duplicate, assessed the full texts of potentially relevant studies for inclusion. Studies were included if they had been conducted in underrepresented populations with rheumatoid arthritis, identified using the PROGRESS-Plus framework. Once we had identified the studies describing the development or implementation of the HAQ, we worked in 2 teams of 2 steering group members to independently extract the relevant data using the existing SOMP table. This permitted us to assess which OMERACT filter criteria had been assessed among different PROGRESS-Plus populations, and draft the Equity Extension, described in Results. Of note, we did not assess the 'domain match' of the HAQ because we assumed that the HAQ has already been proven to match with the content/concept. We did not assess 'clinical trial discrimination' or 'thresholds of meaning' because these need to await the results of relevant studies to be completed.

Table 1.
Original SOMP.

Author/Year	Truth Domain Match	Feasibility	Truth	Discrimination			
			Construct Validity	Test retest Reliability	Longitudinal Construct Validity (Responsiveness)	Clinical Trial Discrimination	Thresholds of Meaning
Lennon 1991			+				
McCartney 2004					+		
Harrison 2004					+	+/-	
Starr 2005				+	+/-	+	+
Best 2006					+		+
Sutcliffe 2006							+
Boers 2007					+/-		+/-
Tugwell 2009							+
Strand 2010	+						
Simon 2010				+	-		+
Brooks 2015	+						
Total available studies for each property	**2**	**0**	**2**	**2**	**6**	**2**	**6**
Total studies available for synthesis	**2**	**0**	**1**	**2**	**6**	**2**	**6**
Rating (RAGW) [put on Master Checklist]	**Green**	**Green**	**Amber**	**Green**	**Green**	**Amber**	**Green**
Overall rating for instrument across properties [Options: Endorsed, Provisional Endorsement, Not endorsed]	**Provisional endorsement: needs additional construct and RCT discrimination**						

Table 2.
SOMP-Equity extension.

PROGRESS Elements	A. Truth Domain match	B. Feasibility*	Truth	Discrimination			
			C. Construct validity	D. Test retest reliability	E. Longitudinal construct validity (responsiveness)	F. Clinical trial discrimination	G. Thresholds of meaning
Place of residence							
Race, culture, language							
Occupation/Employment status							
Gender/sex							
Religion							
Education/literacy							
Socioeconomic status							
Social capital							
Aged (elderly)							

Results

Search results

Results of the literature search can be found in Appendix 1. Our search identified 19,786 records after the removal of duplicates. We excluded studies that did not present complete results (e.g. abstracts), those that did not include a population with rheumatoid arthritis, those that were not assessing the HAQ, and those that did not analyze data across a PROGRESS-Plus characteristic. We included 34 studies assessing the HAQ.

Table 3.
Summary of SOMP-Equity Extension Table completed for HAQ.

Instrument: HAQ PROGRESS Elements	Content/ concept match	Feasibility*	Truth	Discrimination			
			Construct validity	Test retest reliability	Responsiveness	Clinical trial discrimination	Thresholds of meaning
Place of residence	Not assessed	None found	Shakibi 2012	None found	None found	Not assessed	Not assessed
Race, Culture, Language	Not assessed	Chatzitheodorou 2008 Munchey 2018 Citera 2004 Vaidya 2019 Shakibi 2012 Shehab 1998 Abourazzak 2008 Al-Jarallah 1999 Cardiel 1993 Ekdahl 1988 ElMeidany 2003 Esteve-Vives 1993 Kumar 2002 Kirwan 1986 Guillemin 1992	Chatzitheodorou, 2008 Citera 2004 Vaidya 2019 Matsuda 2003 Nazary-Moghadam 2017 Osiri 2001 Osiri 2009 Oude Voshaar 2013 Ranza 1993 Shakibi 2012 Shehab 1998 Song 2014 Tammaru 2007 Thorsen 2001 Abourazzak 2008 Cardiel 1993 Ekdahl 1988 ElMeidany 2003 el-Miedany 2003 Esteve-Vives 1993 Kumar 2002 Koh 1998 Guillemin 2012 Hu 2017 Islam 2013	Chatzitheodorou 2008 Citera 2004 Vaidya 2019 Matsuda 2003 Nazary-Moghadam 2017 Osiri 2009 Ranza 1993 Shakibi 2012 Shehab 1998 Song 2014 Tammaru 2007 Thorsen 2001 Abourazzak 2008 Al Jarallah 1999 Cardiel 1993 Abourazzak 2008 Ekdahl 1988 ElMeidany 2003 Esteve-Vives 1993 Kumar 2002 Koh 1998 Hu 2016 Islam 2013 Ferraz 1990 Guillemin 2012 Hu 2017 Linde 2008	Osiri 2001 Cardiel 1993 el-Miedany 2003 Kumar 2002 Linde 2008	Not assessed	Not assessed
Occupation/ Employment status	Not assessed	None found	Hifinger 2018	None found	None found	Not assessed	Not assessed
Gender/sex	Not assessed	None found	Hifinger 2018 Klooster 2008 Gardiner 1993 Oude Voshaar 2013 Shakibi 2012 Thorsen 2001	None found	None found	Not assessed	Not assessed
Education/ literacy	Not assessed	Citera 2004 Osiri 2009 Tammaru 2007 Thorsen 2001	Citera 2004 Hifinger 2018	None found	None found	Not assessed	Not assessed
Socioeconomic status	Not assessed	Citera 2004	Citera 2004 Shebab 1998	None found	None found	Not assessed	Not assessed
Aged (elderly)	Not assessed	None found	Chatzitheodorou 2008 Munchey 2018 Gardiner 1993 Hifinger 2018 Klooster 2008 Gardiner 1993 Oude Voshaar 2013 Thorsen 2001 Esteve-Vives 1993	None found	None found	Not assessed	Not assessed

Table 4.
Example of the types of information extracted.

PROGRESS Elements	Truth Domain match	Feasibility*	Truth *Evidence of validity of scores*
Occupation/Employment status			**HIFINGER 2018.** For employment status, people who were working had difficulty with different types of items compared to people who were not working (e.g. tasks involving the hands).
Gender/sex			**HIFINGER 2018.** For gender, men had less difficulty with items that were more physically demanding and more with dexterity and women were opposite.
Education/literacy			**HIFINGER 2018.** For education, 10 of 30 items did not capture the same thing for people with more years of education than those with fewer years (however, no clear difference between upper and lower limb activities).
Aged (elderly)			**HIFINGER 2018.** People who were older responded differently to 14 of 30 items compared to people who were younger. For example, older patients had less difficulty with hand function but more difficulty with physically demanding activities.

Development of SOMP Equity Extension

We developed the proposed OMERACT SOMP-Equity Extension shown in Table 2. We retained the Filter 2.1 measurement criteria of the Original SOMP as columns and have listed the equity considerations, using PROGRESS characteristics, as rows. We included all PROGRESS characteristics. For 'Plus', we used 'age' as an important characteristic for arthritis. However, other conditions may choose to include other or additional characteristics, as necessary.

We then took the HAQ-DI example and pilot-tested the fit of available evidence. As mentioned above, we decided that Column A, Domain Match, should have been decided in the Original SOMP (i.e. is generally accepted to match with Domain of Disability). In the future we may wish to explore the degree to which more ideographic methods could help verify if there are any differences in the understanding of a domain (e.g. what is "difficulty cooking a meal" across cultures?) across equity categories. For the purposes of this paper we focused on published literature which would by nature focus on the more traditional measurement properties.

Columns: Our pilot testing of the OMERACT SOMP Equity extension focused on Columns B: '*Feasibility*'; C: '*Construct Validity*'; D: '*Test-retest Reliability*'; and E: '*Responsiveness*'.

For Column B: *'Feasibility'*, we extracted explicit descriptions of patient or provider perspectives as reported in the studies on issues related to access, cost, and time, equipment, or training required.

For Column C: '*Construct validity'*, we extracted descriptions of the population groups included in the studies and whether the measurement instrument was tested across PROGRESS-Plus groups.

For Column D: *'Test-retest reliability'*, we extracted explicit descriptions of the testing of the measurement property across the PROGRESS-Plus groups. For example, comparisons across PROGRESS-Plus linguistic groups within the study or compared to other examples from the literature.

For Column E: '*Responsiveness*' we looked for data reflecting whether there was generation of meaningful thresholds to compare across PROGRESS-Plus groups or subgroups.

Rows: For the virtual meeting we pilot tested the following PROGRESS-Plus characteristics: Race/Culture/Language; Occupation/Employment Status; Sex/Gender; Education/literacy; Socioeconomic Status; Age (elderly) [2].

The information was extracted from the identified studies on the HAQ and entered in the appropriate row and column. This resulted in the information provided in Table 3 (additional details and complete references are available in Appendix 2). Several examples were found where measurement properties were provided across equity groups. As expected, the most commonly reported equity considerations were related to language groups where standards exist for cross cultural adaptation of measurement properties across language versions after an adaptation is completed. Other studies included additional characteristics. For example, the study by Hifinger et al. [10] specifically looked at the performance of the HAQ-DI across several of the PROGRESS-Plus elements including employment status, gender, literacy, and age groups. Examples of the type of information extracted for the Hifinger 2018 study are shown in Table 4.

Discussion

The SOMP-Equity Extension allows researchers to assess whether there are differences in instrument acceptability and performance across sociodemographic characteristics (i.e., PROGRESS-Plus). Studies were found, though several cells remain empty.

Few studies looked at the Filter elements of discrimination, responsiveness and thresholds of meaning, but there was promising work done in construct validity and test-retest reliability across cultures/languages. This suggests that equity of measurement performance is something that has been and can be evaluated. We polled participants about whether this SOMP-Equity extension table is useful for highlighting and tracking equity considerations for OMERACT Core Outcome Measurement Instruments. 100% of the participants agreed with 45% indicating it is very useful and 55% indicating it is moderately useful.

The OMERACT SOMP-Equity extension allows trialists and others to assess and describe how equity has been considered in the development and validation of PROM instruments. This follows other work to extend existing tools to include equity considerations, such as the reporting guidelines for randomized controlled trials (CONSORT-Equity), systematic reviews (PRISMA-Equity), and observational studies (STROBE-Equity). The goal of all of these tools is to improve the reporting of equity considerations to increase the usefulness of the evidence for those making decisions about research, policies, programs, and practice.

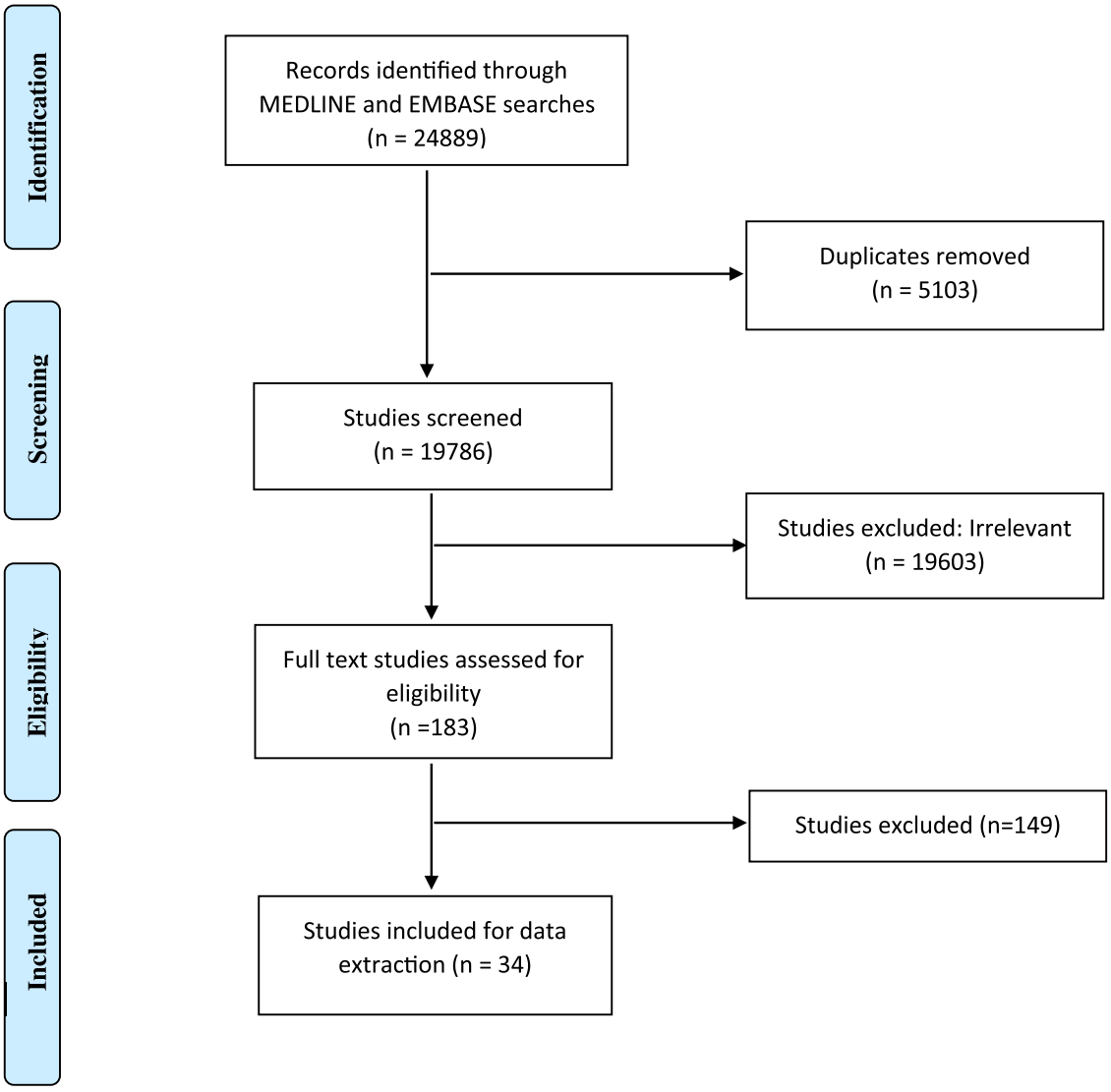
Our next steps will include: (a) a systematic review to assess how equity and population characteristics have been considered in PROMs, in other conditions; this will include looking for evidence on whether these differences affect patients' responses to items; (b) pilot testing this SOMP-Equity Extension using other PROMs in the OMERACT Core Outcome Sets.; (c) developing criteria for rating whether the instrument meets the equity extension for each PROGRESS-Plus characteristic assessed; and (d) initiating discussions with trialists evaluating new PROMs in trials to request inclusion of the SOMP-Equity Extension in new studies.

Acknowledgements

We acknowledge the Equity Working Group members and the participants of the Equity Special Interest Group session: Adewale O Adebajo, Maria Agliotis, Jennifer Barton, Annelies Boonen, Peter Brooks, Rachelle Buchbinder, Flo Callahan, Willemina Campbell, Loreto Carmona, Robin Christensen, Maarten de Wit, Jonathan Epstein, Vicki Evans, Caroline Flurey, Niti Goel, Charlie Goldsmith, Susan Goodman, Rebecca Grainger, Francis Guillemin, Shawna Grosskleg,

Glen Hazlewood, Catherine Hill, Ihsane Hmamouchi, Diana Hollander, Allyson Jones, Janet Jull, Kristin Kelley, Marios Kouloumas, Diane Lacaille, Alex Lee, Anne Lyddiatt, Sabrina Mai Neilsen, Win Min Oo, Jordi Pardo Pardo, Merce Pineros Vegas, Christoph Pohl, Tamara Rader, Sofia Ramiro, Oscar Russell, Nancy Santesso, Grayson Schultz, Jasvinder Singh, Antoine Sreih, Stephanie Taylor, Karine Toupin-April, Courage Uhunmwangho, Vivian Welch, Tiffany Westrich-Robertson.

Appendix 1: Search results PRISMA flow diagram



Appendix 2: SOMP Equity completed HAQ example

Instrument: HAQ			Date completed				
PROGRESS-Plus Elements	Feasibility	Truth Construct validity	Discrimination Test retest reliability	Responsiveness	Clinical trial discrimination	Thresholds of meaning	RATING of meeting the equity extension for this element
Education/ literacy	**Citera 2004.** Excluded Foreign and illiterate patients. **Osiri 2009**. 70% of patients had a limited educational level; no significant variation in the comprehensibility of each item of the Thai HAQ. **Tammaru 2007.** Authors stated that participants clearly comprehended the questions. No description of how this was measured. **Thorsen 2001.** Authors stated that all participants clearly understood the questionnaires. No description of how this was measured.	**Citera 2004.** No correlation with educational level (years of education r=−0.13 p = 0.07). **Hifinger 2018.** "For education, 10 of 30 items showed significant differential item functioning (DIF) but no clear trend could be observed."	No Data	No Data	No Data	No Data	Not assessed
Race, Culture, Language	**Shehab 1998.** (Arabic) Authors state no difficulty with feasibility, objective measurement not provided. provided. **ElMeidany 2003.** (Arabic) All questions were rated as quite or extremely comprehensible (grade 3 and 4). **Abourazzak 2008.** (Moroccan) Authors state no difficulty with feasibility, objective measurement not provided. **Kumar 2002.** (Indian), self-administered with minimal instruction, completion time 3 min **Shakibi 2012.** (Persian) Authors state no difficulty with feasibility, objective measurement not provided. **Munchey 2018.** (Thai) No mention of feasibility. **Vaidya 2019.** (Nepali) Authors state no difficulty with comprehension, objective measurement not provided. **Kirwan 1986.** (British) Authors state no difficulty with feasibility, objective measurement not provided. Two patients required a verbal explanation of how to fill in the questionnaire in addition to the written instructions. **Guillemin 1992.** (French) Authors state no difficulty with feasibility, objective measurement not provided. **Cardiel 1993** (Spanish) Patients were assisted in completing the instrument. **Esteve-Vives 1993.** (Spanish). Completion time 5–10.5 min (mean 6.4) with assistance. 63% of patients could complete the SHAQ in a self-administered way but the remainder 37% could not, mostly due to partial or total illiteracy. **Citera 2004.** (Spanish) Authors state no difficulty with feasibility, objective measurement not provided. **Ekdahl 1988.** (Swedish) No information on feasibility. **Chatzitheodorou 2008.** (Greek) Authors state no difficulty with feasibility, objective measurement not provided.	**Abourazzak 2008.** "Spearman correlation coefficients between all domains comprised between 0.62 and 0.86." **Cardiel 1993.** Convergent and construct validity was obtained for all comparisons (Pearson's $r > 0.4$). **Chatzitheodorou 2008.** Greek version - Most items could be translated with vocabulary equivalence. **Citera 2004.** construct validity showed a good correlation with most of the classic disease activity and functional capacity parameters. **Ekdahl 1988.** No significant differences (Pearson) in ADL Test 2 scores were found between younger and older patients (no sub-group analysis). **ElMeidany 2003.** TJC, RAI, MS, and VAS showed the highest values for Rs in all eight subscales. **el-Miedany 2003.** Significant correlations were found between the ACR response levels and the Arabic-HAQ scores after 6 months (RSpearman 0.438, $P<0.001$) and 12 months (RSpearman 0.594, $P<0.001$). **Esteve-Vives 1993.** cross sectional construct validity, and longitudinal construct validity were similar to other HAQ versions used in different countries. **Kumar 2002.** Construct validity was assessed using Pearson's correlation coefficient between the corresponding values of HAQ and DAS28, both at baseline ($r = 0.49$, $P<0.05$) and after intervention ($r = 0.62$, $P<0.01$). **Koh 1998.** Significant correlation between Chinese HAQ and morning stiffness, tender and swollen joint counts, grip strength, ESR, pain, patient and physician assessment of disease activity. **Guillemin 2012.** Five factors provided by principal components analysis accounted for 75% of the variability of the HAQ score (construct validity). **Hu 2017.** "Construct validity was assessed through evaluating the correlations between HAQ scores and different items and scales in EQ-5D and SF-12″ **Islam 2013:** " Spearman's correlations between the B–HAQ and the other clinical and patient–reported outcome measures were; Pain $r = 0.451$; morning stiffness $r = 0.437$; tender joint count $r = 0.429$; Swollen joint count r0.515; Erythrocyte sedimentation rate $r = 0.258$). **Matsuda 2003**. Japanese HAQ included cultural modifications of 3 questions; high correlation between direct Japanese translation of original HAQ and J-HAQ. **Nazary-Moghadam 2017**. Persian HAQ included cultural modifications of several questions; moderate to strong correlations between all P-HAQ and SF-36 subscales. **Osiri 2001.** Thai HAQ required changes to 4 questions and the addition of 2 activities; the addition of the 2 activities did not change the scoring results. Thai HAQ scores had positive correlation with tender joint count, patient GA, and physician GA. **Osiri 2009**. There was moderate correlation between the majority of Thai HAQ domains, and between the Thai HAQ and disease activity. Highest correlation coefficient was between the Thai HAQ and ACR functional class (CC 0.57); lowest was between the Thai HAQ and ESR (CC 0.37).	**Abourazzak 2008.** Cronbach's alpha showed strong reliability among the 20 items. Test–retest reliability showed a strong reliability with high values for kappa (The kappa test ranged from 0.70 to 0.92 for all domains) and ICC = 0.987. (compared only with study population). **Cardiel 1993.** Reliability, measured by a test-retest with a one-month interval, was high (Spearman's rho = 0.89). (compared only with study population). **Chatzitheodorou 2008.** Assessed concurrent validity of HAQ-GrV against the HADS. (Though they didn't compare validity of Greek vs. English version). **Citera 2004.** it was highly reliable. "Questionnaire reproducibility on day 1 and on day 5 was $r = 0.97$ ($P = 1 \times 10-5$)," (but reproducibility was not compared across PROGRESS+, just within Argentinian version). **Ekdahl 1988.** Inter-observer reliability was high for the ADL (r (S)=0.98), for joint mobility (r(S) =0.86), and for the Ritchie index (r(S)=0.83 (compared only with study population). **ElMeidany 2003**. Cronbach's alpha showed a strong reliability with a standardized alpha of 0.979 among the 20 items (compared only with study population). **Esteve-Vives 1993.** The Pearson correlation coefficient was very good ($r = 0.89$, $p<0.0001$). Compared only with study population. **Kumar 2002.** Iintraclass correlation coefficient: English 0.93, Hindi 0.73. Compared only with study population). **Koh 1998.** Cronbach's alpha was high with value 0.86. Except for walking and grip strength dimensions (compared only with study population). **Hu 2016.** No test-retest reliability. Cronbach's alpha showed a strong reliability. When including the items about the use of aids and devices, Cronbach's alpha was 0.963. When excluding the item about aids and devices, Cronbach's alpha increased to 0.987. Furthermore, this trend was consistent in the subgroups from both patients from north and south China. **Islam 2013.** No test-retest reliability test. Cronbach's alpha was satisfactory for individual-level analyses. **Ferraz 1990.** The test-retest (correlation coefficient =0.905) and Interobserver reliability (correlation coefficient=0.830) was considered satisfactory. Compared only with study population. **Guillemin 2012.** Significant correlation with clinical and radiological variables and reproducible (r intraclass =0.964). **Hu 2017.** Test-retest reliability was not performed. **Linde 2008.** "RAQoL and HAQ	**Cardiel 1993.** "The instrument was sensitive in detecting clinical improvement. Sensitivity to change was 33%, coefficient of responsiveness was −1.04 for improvement. **el-Miedany 2003.** The total Arabic-HAQ index was more sensitive to change after 6 and 12 months. Five of the eight domain subscores had a RE greater than 1 after 12 months; the exceptions were "eating", "hygiene", and "reach". RE in relation to the tender joint count was 1 for "dressing", indicating that sensitivity to change was identical for these two measures. **Kumar 2002.** After treatment, the HAQ values changed to 0.81+/- 0.47 and 0.65+/- 0.55, respectively, demonstrating a very good sensitivity to change (Student's unpaired t-test: $P<0.05$). **Linde 2008.** "SF-36 bodily pain scale and VAS pain were responsive to both improvement and deterioration." **Osiri 2001.** "Thai HAQ scores correlated significantly with some clinical variables after 6 months of DMARD treatment". **Osiri 2009**. Responsiveness of the Thai HAQ was moderate and clinically significant (0.75)– compared values at baseline after 3 months of DMARD treatment.	No Data	No Data	Not assessed

(continued)

(*Continued*)

Instrument: HAQ			Date completed				
PROGRESS-Plus Elements	Feasibility	Truth Construct validity	Discrimination Test retest reliability	Responsiveness	Clinical trial discrimination	Thresholds of meaning	RATING of meeting the equity extension for this element
		Oude Voshaar 2013. Items 3 and 7 were slightly more difficult for US patients than Dutch patients, but the impact of DIF on total HAQ-II scores was negligible, supporting the crosscultural equivalence of Dutch and US HAQ-II scores. **Ranza 1993**. Italian version of HAQ required modification of two questions; close correspondence between ARA functional class (physician-attributed functional status) and HAQ FDI score. **Shakibi 2012**. Persian HAQ included cultural modification of 5 items; high internal consistency between responses to different items (alpha = 0.94), acceptable Spearman's correlation coefficient (r = 0.5) when compared to SF-36 questionnaire scores. **Shehab 1998**. Arabic HAQ required modification of 1 item; all correlations between HAQ scores and disease activity variables were significant. **Song 2014**. MDHAQ-Chinese required several wording modifications; highly strong correlation with English HAQ (coeff of 0.859), moderate to highly strong correlation with all scales of SF-36 (0.528–0.854.) **Tammaru 2007**. Estonian HAQ was able to distinguish between different levels of self-perceived disease severity. **Thorsen 2001.** Dutch HAQ (using scoring method 2, where use of assistive device does not increase score) was able to distinguish between patient-perceived severity, patients rating their day good or bad, and functional status. **Vaidya 2019.** Spearman coefficient for pain and stiffness indicate an adequate construct validity of Nepali HAQ with moderate association with other parameters tested." (Construct validity was examined of the Nepali HAQ, but not across PROGRESS+).	displayed good repeatability (ICC > 0.95) and internal consistency (Cronbach's alpha > 0.90)". **Matsuda 2003.** strong test-retest reliability (not compared across PROGRESS+, just within the Japanese population). **Nazary-Moghadam 2017.** ICC was 0.98–strong test retest reliability (not compared across PROGRESS+, just within Persian population). **Osiri 2009.** ICC was 0.89–acceptable test retest reliability (not compared across PROGRESS+, just within Thai population). **Ranza 1993.** Spearman correlation coefficient was 0.97 (not compared across PROGRESS+, just within Italian population). **Shakibi 2012.** Correlation coefficient was 0.86 (not compared across PROGRESS+, just within Persian population). **Shehab 1998.** Test-retest reliability was 0.81 for overall score, ranged from 0.66 to 0.9 for subscale scores (not compared across PROGRESS+, just within Arabic population). **Song 2014**. Evaluated test-retest reliability but did not report findings (MDHAQ-Chinese). **Tammaru 2007**. The test–retest reliability of the Estonian HAQ-DI was as high as 0.91. **Thorsen 2001.** Dutch HAQ test-retest reliability was 0.90 and 0.93, depending on scoring method. **Vaidya 2019.** "Test–retest reliability of total Nepali HAQ and each item were acceptable. The estimates of ICC for each item ranged from 0.71 to 0.95. The ICC for total Nepali HAQ was 0.763 (CI 0.665 to 0.832)." (Test-retest reliability was examined of the Nepali HAQ, but not across PROGRESS+).				
Aged (elderly)	No Data	**Chatzitheodorou 2008.** Difference in HAQ scores for participants aged <45 years and >45 years. (No further information provided). **Munchey 2018.** Post hoc analyses showed difference in HAQ scores for participants aged 41–60 and >60. **Gardiner 1993.** No significant change in score by age **Hifinger 2018.** Age was related to DIF for 14 of 30 items. "Controlling for overall disability, older patients were less likely to indicate difficulty in performing tasks involving hand function and transfers, and more likely to indicate difficulty for physically demanding activities."(DIF across ages groups and raised a call for more research). **Klooster 2008.** HAQ-DI showed DIF for hygiene (Hygiene was less difficult for younger patients); HAQ-II showed DIF for getting on & off toilet, standing up from a straight chair, and opening car doors (all more difficult for younger patients) (DIF across ages groups and raised a call for more research). **Gardiner 1993.** Age was significantly related to baseline HAQ score, but not to change in HAQ score after 5 years (no discussion on whether this was due to disease severity vs. questionnaire bias). **Oude Voshaar 2013.** Found all HAQ-II items functioned equivalently across age (used DIF analysis on combined Dutch & US samples, created 3 equally large age groups). **Thorsen 2001.** Scores on the Dutch HAQ were not associated with age. **Esteve-Vives 1993.** Correlation found between age and SHAQ scores but no sub-group analysis carried out.	No Data	No Data	No Data	No Data	Not assessed
Employment status/ Occupation	No Data	**Hifinger 2018.** For employment status, significant DIF was seen in 19 of 30 items. "Controlling for overall disability, patients who were in paid or unpaid employment were more likely to report difficulties with tasks involving the hands but less likely to report difficulty with more strenuous activities involving lower limb function."	No Data	No Data	No Data	No Data	Not assessed

(continued)

(*Continued*)

Instrument: HAQ			Date completed				
PROGRESS-Plus Elements	Feasibility	Truth Construct validity	Discrimination Test retest reliability	Responsiveness	Clinical trial discrimination	Thresholds of meaning	RATING of meeting the equity extension for this element
Sex/ Gender	No Data	**Gardiner 1993.** Sex was not significantly related to baseline HAQ score, but not to change in HAQ score after 5 years (discussion references another study that suggests differences in HAQ between sexes are due to disease severity rather than questionnaire bias). **Hifinger 2018.** For gender, "significant DIF was observed in 23 of 30 HAQ items. Compared to males with the same overall disability, females reported systematically less difficulties for items related to dressing and grooming as well arising, whereas they reported more difficulties for items that require hand strength or are physically more demanding." **Klooster 2008.** HAQ-DI showed DIF for dressing (women had less difficulty) and grip (women had more difficulty); HAQ-II showed DIF for standing up from a straight chair (more difficult for men). **Oude Voshaar 2013.** Found all HAQ-II items functioned equivalently across sex (used DIF analysis on combined Dutch & US samples). **Shakibi 2012.** Persian HAQ had high internal consistency for both females (alpha = 0.94) and males (alpha = 0.94). When comparing scores to the SF-36, males (r = 0.63) had a higher correlation than females (r = 0.49). **Thorsen 2001.** Scores on the Danish HAQ were not associated with gender.	No Data	No Data	No Data	No Data	Not assessed
Socioeconomic status	**Citera 2004.** Authors stated that there was no difficulty in completing the questionnaire (no numerical data).	**Citera 2004.** "A weak although significant negative correlation was found between the HAQ-A and the economic level (measured as average monthly income; r=−0.21 P = 0.03)" (Did not discuss whether this was due to income affecting how patients use the tool or income affecting the severity of RA). **Shebab 1998.** 23% of women answered 'not applicable' to the item 'do chores such as vacuuming or yardwork' because they had domestic employees and "may warrant changing the item for another activity more relevant to Kuwaiti culture".	No Data	No Data	No Data	No Data	Not assessed
Place of residence	No Data	**Shakibi 2012.** Persian HAQ had high internal consistency for both urban residents (alpha=0.93) and rural residents (alpha=0.97). When comparing scores to the SF-36, urban citizens (r = 0.57) had a higher correlation than rural citizens (r = 0.42).	No Data	No Data	No Data	No Data	Not assessed

'Feasibility' refers to consideration of the feasibility across PROGRESS+ group, such as access, costs, equipment required, training needed, burden, etc.

'Truth' and 'construct validity' refer to whether the measurement instrument has been tested across the PROGRESS+ groups

'Test-retest reliability' refers to whether the measurement instrument has been tested across the PROGRESS+ groups or across literature for the different groups

By 'responsiveness' we are assessing whether the instrument can detect changes over time across PROGRESS+ groups

References for included studies

Abourazzak FE, Benbouazza K, Amine B, Bahiri R, Lazrak N, Bzami F, et al. Psychometric evaluation of a Moroccan version of health Assessment questionnaire for use in Moroccan patients with rheumatoid arthritis. Rheumatology International. 2008;28(12):1197–203.

Al-Jarallah K, Shehab D, Al-Saeid K, Moussa MAA. Measurement of the functional status in juvenile rheumatoid arthritis: Evaluation of the Arabic version of the childhood health assessment questionnaire. Medical Principles and Practice. 1999;8(4):281–6.

Cardiel MH, Abello-Banfi M, Ruiz-Mercado R, Alarcon-Segovia D. How to measure health status in rheumatoid arthritis in non-English speaking patients: validation of a Spanish version of the Health Assessment Questionnaire Disability Index (Spanish HAQ-DI). Clinical and experimental rheumatology. 1993;11(2):117–21.

Chatzitheodorou D, Kabitsis C, Papadopoulos NG, Galanopoulou V. Assessing disability in patients with RHEUMATIC diseases: Translation, reliability and validity testing of a Greek version of the Stanford health Assessment Questionnaire (HAQ). Rheumatology International. 2008;28 (11):1091–7.

Citera G, Arriola MS, Maldonado-Cocco JA, Rosemffet MG, Sánchez MM, Goñi MA, et al. Validation and crosscultural adaptation of an argentine spanish version of the health assessment questionnaire disability index. JCR: Journal of Clinical Rheumatology. 2004;10(3):110–5.

Ekdahl C, Eberhardt K, Andersson SI, Svensson B. Assessing disability in patients with rheumatoid arthritis. Scandinavian Journal of Rheumatology. 1988;17(4):263–71.

Esteve-Vives J, Batlle-Gualda E, Reig A, Tornero J, Tenorio M, Nunez M. Spanish version of the Health Assessment Questionnaire: Reliability, validity and transcultural equivalency. Canada Journal of Rheumatology. 1993;20(12):2116–22.

Ferraz M, Oliveira LM, Araujo PM, Atra E, Tugwell P. Crosscultural reliability of the physical ability dimension of the health assessment questionnaire. The Journal of rheumatology. 1990;17(6):813–7.

Gardiner PV, Sykes HR, Hassey GA, Walker DJ. An evaluation of the health assessment questionnaire in long-term longitudinal follow-up of disability in rheumatoid arthritis. Rheumatology. 1993;32(8):724–8.

Guillemin F, Briançon S, Pourel J. Validity and discriminant ability of the HAQ functional index in EARLY rheumatoid arthritis. Disability and Rehabilitation. 1992;14(2):71–7.

Hifinger M, Norton S, Ramiro S, Putrik P, Sokka-Isler T, Boonen A. Equivalence in the health Assessment questionnaire (haq) Across socio-demographic determinants: ANALYSES within quest-ra. Seminars in Arthritis and Rheumatism. 2018;47(4):492–500.

Hu H, Luan L, Yang K, Li S-C. Psychometric validation of Chinese health Assessment questionnaire for use in rheumatoid arthritis patients in China. International Journal of Rheumatic Diseases. 2016;20(12):1987–92.

Islam N, Baron Basak T, OudeVoshaar MA, Ferdous N, Rasker JJ, Atiqul Haq S. Cross-cultural adaptation and validation of a Bengali health Assessment questionnaire for use IN rheumatoid arthritis patients. International Journal of Rheumatic Diseases. 2013;16(4):413–7.

Kirwan JR, Reeback JS. STANFORD health assessment questionnaire modified to Assess disability in British patients with rheumatoid arthritis. Rheumatology. 1986;25(2):206–9.

Klooster PM, Taal E, van de Laar MAFJ. Rasch analysis of the Dutch Health Assessment Questionnaire disability index and the Health Assessment Questionnaire II in patients with rheumatoid arthritis. Arthritis and Rheumatism. 2008;59 (12):1721–8.

Koh ET, Seow A, Pong LY, Koh WH, Chan L, Howe HS, et al. Cross cultural adaptation and validation of the Chinese Health Assessment Questionnaire for use in rheumatoid arthritis. The Journal of rheumatology. 1998;25(9):1705–9.

Kumar A. Validation of an Indian version of the health Assessment questionnaire in patients with rheumatoid arthritis. Rheumatology. 2002;41(12):1457–9.

Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland M. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D [corrected] RAQoL, and HAQ in patients with rheumatoid arthritis. The Journal of rheumatology. 2008;35(8):1528–37.

Matsuda Y, Singh G, Yamanaka H, Tanaka E, Urano W, Taniguchi A, et al. Validation of a Japanese version of the Stanford health Assessment questionnaire In 3763 patients with rheumatoid arthritis. Arthritis & Rheumatism. 2003;49(6):784–8.

Meidany YME, Gaafary MME, Ahmed I. Cross-cultural adaptation and validation of an arabic health assessment questionnaire for use in rheumatoid arthritis patients. Joint Bone Spine. 2003;70(3):195–202.

Miedany YE, Youssef S, Gaafary ME, Ahmed I. Evaluating changes in health status: Sensitivity to change of the MODIFIED Arabic health assessment questionnaire in patients with rheumatoid arthritis. Joint Bone Spine. 2003;70(6):509–14.

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 in Health Regional Issues. 2018;15:76–81.

Nazary-Moghadam S, Zeinalzadeh A, Salavati M, Almasi S, Negahban H. Adaptation, reliability and validity testing of a Persian version of the health Assessment questionnaire-disability index in Iranian patients with rheumatoid arthritis. Journal of Bodywork and Movement Therapies. 2017;21(1):133–40.

Osiri M, Wongchinsri J, Ukritchon S, Hanvivadhanakul P, Kasitanon N, Siripaitoon B. Comprehensibility, reliability, validity, and responsiveness of the Thai version of the health Assessment questionnaire in Thai patients with rheumatoid arthritis. Arthritis Research & Therapy. 2009;11(4).

Osiri M. Evaluation of functional ability of Thai patients with rheumatoid Arthritis by the use of a Thai version of the health Assessment Questionnaire. Rheumatology. 2001;40(5):555–8.

Oude Voshaar MA, Glas CA, ten Klooster PM, Taal E, Wolfe F, van de Laar MA. Crosscultural measurement equivalence of the health assessment questionnaire ii. Arthritis Care & Research. 2013;65(6):1000–4.

Ranza R, Marchesoni A, Calori G, Bianchi G, Braga M, Canazza S, et al. The Italian version of the Functional Disability Index of the Health Assessment Questionnaire. A reliable instrument for multicenter studies on rheumatoid arthritis. Clinical and experimental rheumatology. 1993;11(2):123–8.

Shakibi MR. Validation of the personal impact health Assessment questionnaire in patients with rheumatoid arthritis in Kerman, iran. Turkish Journal of Rheumatology. 2012;27(2):121–7.

Shehab D, Al-Jarallah K, Moussa MA. Validation of the Arabic version of the Health Assessment Questionnaire (HAQ) in patients with rheumatoid arthritis. Revue du rhumatisme. 1998;65(6):387–92.

Song Y, Zhu L-an, Wang S-li, Leng L, Bucala R, Lu L-J. Multi-Dimensional health Assessment questionnaire in CHINA: RELIABILITY, validity and clinical value in patients with rheumatoid arthritis. PLoS ONE. 2014;9(5).

Tammaru M, Singh G, Hanson E, Maimets K. The adaptation, reliability and validity testing of the Estonian version of the health Assessment questionnaire's Disability Index. Rheumatology International. 2007;28(1):51–9.

Thorsen H, Hansen TM, McKenna SP, Sorensen SF, Whalley D. Adaptation into Danish of the Stanford health Assessment questionnaire (HAQ) and the rheumatoid arthritis quality of life scale (RAQOL). Scandinavian Journal of Rheumatology. 2001;30(2):103–9.

Vaidya B, Joshi R, Lama LD, Nakarmi S. Translation, cross–cultural adaptation and validation of Nepali version of health Assessment questionnaire–disability index IN rheumatoid arthritis patients from Nepal. International Journal of Rheumatic Diseases. 2019;22(10):1871–6.

References

[1] Whitehead M. The concepts and principles of equity and health. Int J Health Serv 1992;22(3):429–45.

[2] O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, 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(1):56–64.

[3] Petkovic J, Epstein J, Buchbinder R, Welch V, Rader T, Lyddiatt A, et al. Toward ensuring health equity: readability and cultural equivalence of OMERACT patient-reported outcome measures. J Rheumatol 2015;42(12):2448–59.

[4] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67(7):745–53.

[5] Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham CO, et al. OMERACT Filter 2.1: elaboration of the conceptual framework for outcome measurement in health intervention studies. J Rheumatol 2019;46(8):1021–7.

[6] Beaton DE, Maxwell LJ, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Instrument selection using the OMERACT Filter 2.1: the OMERACT methodology. J Rheumatol 2019;46(8):1028–35.

[7] Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. J Rheumatol 1982;9(5):789–93.

[8] Bruce B, Fries JF. The health assessment questionnaire (HAQ). Clin Exp Rheumatol 2005;23(5 Suppl 39):S14–8.

[9] Ramey DR, Raynauld JP, Fries JF. The health assessment questionnaire 1992: status and review. Arthritis Care Res 1992;5(3):119–29.

[10] Hifinger M, Norton S, Ramiro S, Putrik P, Sokka-Isler T, Boonen A. Equivalence in the Health Assessment Questionnaire (HAQ) across socio-demographic determinants: analyses within QUEST-RA. Semin Arthritis Rheum 2018;47(4):492–500.