~~HAQ-DI and the SF-36 Physical Functioning subscale provisionally endorsed for the measurement of the physical function domain in psoriatic arthritis using OMERACT methodology~~

OMERACT Filter 2.1 Instrument Selection for Physical Function domain in psoriatic arthritis: Provisional Endorsement for HAQ-DI and SF-36 PF.

Ying Ying Leung[1,2], Ana-Maria Orbai[3], Pil Hojgaard[4, 5], Richard Holland[6], Ashish J Mathew[7,8,9], Niti Goel[10], Jeffrey Chau[11], William Tillett[12], Christine Lindsay[13], Alexis Ogdie[14], Laura C Coates[15], Dafna D Gladman[16], Robin Christensen[17], Philip Mease[18], Vibeke Strand[19]

1. Singapore General Hospital, Duke-NUS Medical School, Singapore
2. Duke-NUS Medical School, Singapore
3. Division of Rheumatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
4. Department of Rheumatology, Holbaek Hospital, Denmark
5. Musculoskeletal Statistics Unit, The Parker Institute, Denmark
6. Concord Repatriation General Hospital, Sydney, Australia
7. Centre for Prognosis Studies in Rheumatic Diseases, Division of Rheumatology, Department of Medicine, University of Toronto, Toronto, Ontario Canada
8. Copenhagen Center for Arthritis Research (COPECARE), Rigshospitalet Glostrup, University of Copenhagen, Copenhagen, Denmark
9. Department of Clinical Immunology & Rheumatology, Christian Medical College, Vellore, India
10. Patient Research Partner, Adjunct Assistant Professor, Duke University School of Medicine, Durham, North Carolina, USA
11. Patient Research Partner, Hong Kong

12. Royal National Hospital for Rheumatic Diseases, University of Bath, Bath, United Kingdom
13. Patient research partner. Employed by Aurinia Pharma US Inc., Prosper, Texas USA
14. Medicine and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
15. National Institute for Health Research Clinician Scientist, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom
16. Medicine, University of Toronto, Senior Scientist, Krembil Research Institute, Director, Psoriatic Arthritis Program, University Health Network, Toronto Western Hospital, Toronto, Ontario, Canada
17. Bispebjerg and Frederiksberg Hospital, Copenhagen & Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark.
18. Rheumatology Research, Swedish Medical Center and University of Washington School of Medicine, Seattle, Washington, USA
19. Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto, California, USA

Correspondence to: Ying-Ying Leung, MD; Department of Rheumatology and Immunology, Singapore General Hospital, The Academia, level 4, 20 College Road, Singapore 169856, Contact No.: +65 63265276, Fax no.: +65 62203321, E-mail: katyccc@hotmail.com

Running Title: HAQ-DI and SF-36 PF endorsed using OMERACT methodology

Abstract

Objectives. Physical function is one of the core domains to be measured in all trials in psoriatic arthritis (PsA). We aimed to evaluate two instruments for physical function in PsA: The Health Assessment Questionnaire-disability index (HAQ-DI) and the physical functioning subscale of the Medical Outcome Survey Short-Form 36 items (SF-36 PF).

Methods. We followed guidelines set out by the OMERACT Filter 2.1. A working group was formed to evaluate each instrument for domain match and feasibility to reach consensus. Two systematic literature reviews (SLRs) were conducted to identify the relevant articles supporting measurement properties of both instruments. Five additional measurement properties were appraised: construct validity, test-retest reliability, longitudinal construct validity, clinical trial discrimination, and threshold of meaning. New evidence was synthesized to fill the gap. Data were presented to the OMERACT technical advisory group (TAG) and the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) community for endorsement.

Results. The results for seven measurement properties for HAQ-DI and SF-36 PF were presented in Summary of Measurement Property (SOMP) tables. The working group proposed “Provisional Endorsement” for both instruments. The body of evidence was approved by the OMERACT TAG. In two Delphi exercises among GRAPPA members, HAQ-DI received 93.9% and 97.5% endorsement votes, while that for SF-36 PF were 86.7% and 77.3%.

Conclusion. Both HAQ-DI and SF-36 PF were provisionally endorsed for the measurement of physical function in PsA trials, using the OMERACT Filter 2.1.

Clinical significance

- HAQ-DI and SF-36 PF have the necessary measurement properties for the measurement of physical function in PsA to be provisionally endorsed by the OMERACT Filter 2.1. This provisional endorsement has provided justification for the use and reporting of HAQ-DI and SF-36 PF to assess physical function in both randomized controlled trials and longitudinal observational studies in PsA.
- These instruments have achieved consensus (i.e., >70% endorsement votes) from the GRAPPA community.
- This effort demonstrated the ongoing commitment of GRAPPA to standardize the core outcome measurement set for PsA.
- This work showed the relevance and feasibility of implementing a data-driven, evidence-based process laid out in the OMERACT Filter 2.1 for instrument selection.

1. Introduction

Psoriatic arthritis (PsA) is a chronic systemic disease with multiple musculoskeletal manifestations including synovitis, dactylitis, enthesitis, and spondylitis (1). However, the impact of PsA extends beyond bone and joint to physical disability, fatigue, and depression. Physical function is also considered one of the core aspects affected by the disease per the patient perspective (2, 3). The Outcome Measures Working Group from the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) updated the core domain set for PsA in 2016, and physical function is among one of the core domains to be measured in all randomized controlled trials (RCTs) and longitudinal observational studies (LOS) (3). After identifying the core domain set, GRAPPA has committed to work towards standardization of clinical trial assessments by developing a core outcome measurement set for each important domain (4, 5) in collaboration with the Outcome Measures in Rheumatology (OMERACT). A standardized core outcome measurement set will minimize the variability in outcomes in RCTs and facilitate robust measurement and comparison across trials. Given its clinical relevance to patients and in clinical trials, physical function is one of the domains that GRAPPA has prioritized for identification of outcomes to include in the core measurement set (5, 6).

OMERACT has published the OMERACT Filter 2.1 Instrument Selection Algorithm (OFISA) to guide the selection of instruments (7). It is based on the three pillars of OMERACT (8): Truth, Discrimination and Feasibility, and it recommends detailed evaluation of each measurement via four signaling questions: matching with target domain (Truth), practicality to use (Feasibility), making numeric sense (Truth), and discriminating between groups of interest (Discrimination). Each instrument needs to be appraised for seven measurement properties to answer one single question: whether there is adequate evidence to support the use of the specific instrument in clinical research, e.g., RCTs or LOS. For each

instrument, the appraisal process starts with a working group ensuring domain match and feasibility. For suitable instruments, measurement properties will subsequently be appraised via systematic literature review. Where there is no existing evidence for a measurement property, new studies will have to be performed to bridge the gap.

Our aim was to document the process and evidence derived to support two instruments for physical function in PsA: The Health Assessment Questionnaire-Disability Index (HAQ-DI) and the physical functioning subscale of the Medical Outcome Survey Short-Form 36 items (SF-36 PF). These two instruments were chosen as most of the evidence supporting the seven measurement properties of the OMERACT Filter 2.1 was available.

2. Methods

We followed the guidelines set out by the OMERACT Filter 2.1 (7). The processes started with convening a working group with the relevant stakeholders inclusive of care providers and patients. The group set to define the research topic and the instruments in focus. The working group discussed and evaluated each instrument for domain match and feasibility. The perspectives from a larger group of patients were subsequently sought. Instruments judged to match the domain and practical to use were further appraised. Systematic literature reviews (SLRs) were conducted in accordance to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) recommendations (9) to identify the relevant articles supporting measurement properties of the instrument in test. In addition to domain match and feasibility, five more measurement properties were appraised: construct validity, test-retest reliability, longitudinal construct validity, clinical trial discrimination, and threshold of meaning (7). In circumstances for which inadequate or no data were available for a certain measurement property, members of the working team designed new research proposals and synthesized new evidence ~~was synthesized~~ to fill the gap.

2.1 Quality appraisal

For each article included in the SLR, at least two working group members independently assessed the quality using the OMERACT Good Method Checklist for each of the five measurement properties. In case of dispute, additional members from the working group were engaged until consensus was reached. In the OMERACT Good Method Checklist, several aspects were assessed for each measurement property and rated as whether good methods were used (yes/no), and a final rating for each measurement property was given as Green (Yes, likely low risk of bias), Amber (Some cautions, but can be used as evidence), and Red (No, don't use this evidence). For each measurement property, articles rated as Red were not included in the subsequent evidence synthesis.

2.2 Adequacy of performance

For each article included in the SLR, the adequacy of performance for each measurement property was assessed by at least two working group members independently; additional members were engaged to resolve disputes. The adequacy of performance for each property was rated as either (+) adequate; (+/-) equivocal; or (-) inadequate. A summary of provisional standard of adequate performance according to OMERACT filter 2.1 is given in the Appendix. The detailed rationale of assessment was documented in table format.

2.3 Final rating for each measurement properties

The evidence supporting the seven measurement properties for each instrument are presented in the summary of measurement property (SOMP) table. The final ratings for each of the seven measurement properties were classified as: GREEN (good to go), AMBER (some cautions, but still can be used), RED (stop, do not use this), or WHITE (no data). This final rating was synthesized in accordance to the OMERACT Filter 2.1 recommendation, taking into account the number of good quality articles available (rated as Green/Amber/Red), the adequacy of the measurement property (rated as +, +/-, and -), and

the consistency across articles (8). In brief, a measurement property supported by at least two good quality articles showing consistent findings with adequate performance was considered GREEN. A measurement property having at least two articles, but with inconsistent findings; or having only one article with inadequate performance was given a RED. In all other situations, a final rating of AMBER was given. A final rating for each outcome measure across the seven measurement properties was generated to address whether it had adequate evidence to support its use in clinical research. This overall rating is reported as "Fully endorsed", "Provisionally endorsed", or "Not endorsed".

2.4 Endorsement of instruments

All the processes and evidence leading to the SOMP tables and overall ratings were summarized in the OMERACT Instrument Selection Workbook for each instrument, which were read by the OMERACT TAG with at least two members from OMERACT. Discussion, clarifications, and verifications were made between the working group and TAG to achieve an endorsement from the TAG.

The body of evidence to support both instruments was presented to the GRAPPA community at the GRAPPA annual scientific meeting in July 2020. Online polling and a subsequent online Delphi exercise were conducted to seek agreement from the GRAPPA community. In both Delphi exercises, GRAPPA members were asked: "Given the evidence, do you agree with the working group to the (overall rating) for the outcome measure (HAQ-DI or SF-36 PF) as core instruments for the measurement of physical function in PsA studies?" A voting of >70% was taken as agreement.

3. Results

3.1 The physical function working group

A working group for physical function domain was developed in June 2018 with 13 members including 2 patient research partners (PRPs) (6) and was expanded to 15 members in July 2019. These working group members came from three continents; and were invited to participate from the GRAPPA community on the basis of experience in physical function measurement in PsA or expressed interest to work on the topic.

Based on the patient perspective from an international qualitative study (3, 10), physical function in PsA is defined as being able to perform physical activities (includes upper/lower extremity functioning). Based on the concept of physical function being the perception of physical capability, the working group therefore decided to focus on patient-reported outcome measures (PROMs) instead of performance-based assessments. The working group discussed and agreed to focus on six PROMs for physical function. The detailed rationale for this selection was published previously (11). The current report is limited to the appraisal for HAQ-DI and SF-36 PF, with the full instrument selection workbooks for both PROMs available on the OMERACT website (OMERACT Home).

3.2 Raw data review

Raw data for HAQ-DI and SF-36 PF were reviewed. A high floor effect of 24.5% was noted for HAQ-DI in an observational study, which reflected data from patients of which 68.4% were in a minimal disease activity state and 14.9% with higher disease activity. The floor effect for SF-36 PF was less remarkable (7.7%).

3.3 Domain match and feasibility

Domain match and feasibility involve the opinion of stakeholders. The 13 working group members discussed the match of HAQ-DI and SF-36 PF to the domain and feasibility in two webinars. We subsequently conducted a Delphi exercise using seven questions evaluating domain match and seven questions, feasibility according to the OMERACT Filter 2.1. The results are summarized in Table 1.

Opinion from patients was sought from a broader group. To engage patients, a video explaining the OMERACT Filter 2.1 methodology and describing information regarding the physical function PROMs was co-developed by the team lead and 2 PRPs in the working group (NG and JC) [https://youtu.be/Qd86PwzgvQI]. All the GRAPPA PRPs then participated in a Delphi exercise for domain match and feasibility after watching the video (Table 1).

Considering the results from both Delphi exercises from working groups and PRPs, the working group consensus on domain match and feasibility were AMBER and GREEN, respectively, for HAQ-DI, and both AMBER for SF-36 PF.

3.4 Systematic literature review (SLR)

For the other five measurement properties, two SLRs were conducted for appraisal. The first SLR included articles with a primary aim to evaluate measurement properties of all PROMs in PsA from observational trials (PROSPERO ID: CRD42016032546) (12). As the first SLR did not include RCTs, the working group conducted the second SLR to identify physical function PROMs from RCTs for the appraisal of clinical trial discrimination (PROSPERO ID: CRD42019129557) (13). The first SLR identified 10 relevant articles for HAQ-DI and six articles for SF-36 PF. Subsequently, new evidence was generated by the working group members to bridge the knowledge gaps for HAQ-DI (n=4 studies) and SF-36 PF (n=1 study). For the second SLR, relevant RCTs were reviewed for data on HAQ-DI (n=31) and for SF-36 PF (n=4). The results of the SLRs are summarized in the SOMP tables for HAQ-DI (Table 2) and SF-36 PF (Table 3).

3.4.1 Quality Appraisal

We used the OMERACT Good Method Checklist for quality appraisal of each article identified from the SLRs for five measurement properties (7, 8). Disputes were reconciled by discussion, and more group members were engaged when needed. Detailed comments for

ratings (Green, Amber, Red, White) for each article were documented in the Instrument selection workbooks.

3.4.2 Evaluation of performance for measurement properties for each outcome measure

For each article and measurement property, the performance of HAQ-DI or SF-36 PF was appraised by at least two group members independently; disputes were reconciled with additional group members.

3.5 Final ratings for measurement properties

3.5.1 Construct validity

For HAQ-DI, eight articles were available for appraisal of construct validity, six were identified from the SLR (14-19), while two were newly developed by the working group members and their respective teams (20, 21). Most articles did not describe the measurement properties of the comparator outcome measures in detail. However, following discussion with working group members and OMERACT TAG, the working group considered this as a minor concern that would not cause a risk of severe bias. Five articles were rated as (+) adequate for construct validity, affirmed by fulfilling >75% of hypothesis set *a priori* for correlations with comparison outcome measures or distinguishing between known groups. Three articles were rated for (+/-) equivocal, with the main concerns either that the *a priori hypothesis* for construct validity did not match with a current standard, or <75% of *a priori* hypothesis were met. The detailed reporting of evidence supporting each measurement property is summarized for HAQ-DI in Appendix Table A.1. Given eight articles of good quality, five rated as having adequate performance, the final rating for construct validity for HAQ-DI was GREEN (Table 2).

For SF-36 PF, four articles were identified in the SLRs for evaluation (16, 17, 22, 23). Two articles were rated Amber for quality due to inadequate description of SF-36 PF or comparator outcome measures. All four articles were rated as (+) adequate for construct

validity (Appendix Table B.1). The final rating for construct validity for SF-36 PF was GREEN (Table 3).

3.5.2 Test-retest reliability

There were no published data on test-retest reliability for HAQ-DI and SF-36 PF in PsA at the beginning of the project. New data and analysis were conducted for test-retest reliability (24). In brief, the~~The~~ working group members identified ~~unpublished~~ data from two studies for HAQ-DI and one for SF-36 PF. Detail of these two studies was published elsewhere and summarized in Appendix A.2 and B.2. In brief, both studies recruited a subset of PsA patients with stable disease without medication change for test-retest reliability. Given PsA is a chronic illness, stability between two short assessment time points can be assumed in patients without medication change. Both studies were rated as Green, low risk of bias and included for evidence synthesis. Both HAQ-DI and SF-36 PF were rated as (+) adequate performance ~~(Appendix A.2 and B.2)~~. The rating for test-retest reliability for HAQ-DI was GREEN (Table 2), while that for SF-36 PF was AMBER as there was only a single dataset available (Table 3).

3.5.3 Longitudinal construct validity

Longitudinal construct validity addresses whether instruments can discriminate between situations of interest. To achieve endorsement requires the instrument to demonstrate a capacity to change with a desirable effect size and a direction of a definite change in the patients' condition with time, usually anchored to external measurement that is meaningful to patients. For SF-36 PF, the quality of the two articles identified from the SLR was rated as Amber due to the small sample sizes in the subgroup of patients who achieved a change in condition (25, 26). With these two articles, longitudinal construct validity was rated as AMBER for SF-36 PF (Table 3). As for HAQ-DI, in addition to the above-mentioned articles, two newly published articles (20, 21) have good quality and showed adequate

performance (Appendix Table A.3). The final rating for longitudinal construct validity was GREEN for HAQ-DI (Table 2).

3.5.4 Clinical trial discrimination

Clinical trial discrimination addresses the degree to which instruments are sensitive to change between the intervention arms in RCTs. A separate SLR was conducted to identify all the RCTs in PsA, and 31 and 4 articles reported results for HAQ-DI and SF-36 PF, respectively. Detailed quality appraisal and adequacy of performance have been published elsewhere (13). For HAQ-DI, 29 articles with good quality affirmed a final rating of GREEN for clinical trial discrimination (Appendix Table A.4). For SF-36 PF, a final rating of AMBER was given, with two articles having some quality concerns (Appendix Table B.4).

3.5.5 Threshold of meaning

Threshold of meaning is the degree to which one can assign easily understood cut-offs to the instruments for interpretations, including minimally important difference (MCID) or patient acceptable symptom state. Three articles identified from the SLR described MCID for HAQ-DI in PsA (26-28). Quality appraisal was rated as Amber as thresholds were not derived from multiple criteria and analysis, and results were not triangulated. Nonetheless, the reported MCID for improvement across studies were consistent, ranging from -0.27 to -0.36 (Appendix Table A.5). There was one newly published article which fulfilled the quality assessment and showed adequacy of performance (20) (Table 2). The final rating for these four articles was AMBER. For SF-36 PF there was only a single article with small sample size (26) (Appendix Table B.5), and the final rating for threshold of meaning was AMBER (Table 3).

3.6 Overall ratings

The results of evidence to support measurement properties for HAQ-DI and SF-36 PF are presented in Tables 2 and 3. Weighing across the seven measurement properties, the

working group proposed overall ratings for HAQ-DI and SF-36 PF as “Provisional endorsement”. Further work will be required for test-retest reliability, clinical trial discrimination and threshold of meaning, particularly for SF-36 PF.

3.7 Endorsement

The instrument selection workbooks that contain the full body of evidence to support measurement properties were submitted to and critically reviewed by the OMERACT TAG. After clarifications and amendments, the OMERACT TAG approved the body of evidence for both instruments in September 2020.

These data were presented to the GRAPPA community during the GRAPPA annual meeting in July 2020, followed by a live polling. Two questions were asked: “Given the evidence, do you agree with the working group to endorse HAQ-DI as provisional core instruments for the measurement of physical function in PsA studies?” and the same question was asked for the SF-36 PF. Forty-nine GRAPPA members participated in the live polling, with 93.9% and 86.7% voting positively for HAQ-DI and SF-36 PF, respectively. A subsequent online Delphi exercise for all GRAPPA members was conducted from 20 October 2020 to 16 November 2020. The original data, updated evidence, and results of the live poll at the GRAPPA 2020 annual meeting were presented. Identical voting questions were used in both surveys. A total of 119 GRAPPA members participated and voted for provisional endorsement of HAQ-DI (97.5% positive) and SF-36 (77.3% positive) as core instruments for the measurement of physical function in PsA studies.

4. Discussion

In this report, we summarize the evidence leading to provisional endorsement of HAQ-DI and SF-36 PF for the measurement of physical function in PsA. These instruments achieved consensus (i.e., >70% endorsement votes) from the GRAPPA community. This

effort demonstrates the ongoing commitment of GRAPPA to standardize the core outcome measurement set for PsA. This work also showed the feasibility of implementing a data-driven, evidence-based process laid out in the OMERACT Filter 2.1 for instrument selection and provides a model for instrument selection for other domains. This effort follows the success in achieving full endorsement of the 66/68 swollen and tender joint count and provisional endorsement of the Psoriatic Arthritis Impact of Disease (PsAID) questionnaire for the assessment of musculoskeletal disease activity and health-related quality of life in PsA, respectively (7, 29, 30). Physical function is one of the core domains to be measured in all RCTs and LOS, while HAQ-DI and SF-36 are the most commonly used instruments in RCTs. This provisional endorsement has provided justification for their continued use and reporting in both RCTs and LOS.

The strength of the current study included comprehensive assessment of the seven measurement properties using a rigorous methodology framework layout by the OMERACT Filter 2.1. It is a combined effort of an international team, with inclusion of PRPs. Throughout the whole process, there was adequate engagement of PRPs who not only provided valuable input to domain match and feasibility, but also developed appropriate education materials for methodology issues for other PRPs, appraised the quality of articles, and assessed the adequacy of measurement properties.

A research agenda has been developed to attain full endorsement for both the HAQ-DI and SF-36 PF. The research agenda includes establishing thresholds of meaning for both instruments in PsA. Four articles evaluated MCID for HAQ-DI in PsA, with three older studies that have either used single anchor or single statistical methods without triangulation. There has been no gold standard for an anchor that constitutes a true change or status of wellness (31). It has been conceptualized that the anchors should be clinically meaningful to patients rather than based on statistical methods. Further studies using the existing RCT

dataset to evaluate several clinically relevant thresholds to stakeholders, using multiple statistical methods for triangulation, are planned. As for clinical trial discrimination, there were only a few trials that reported the results for SF-36 PF despite most RCTs incorporating SF-36 as an outcome measure. Most RCTs have reported the physical function summary scores (PCS) of SF-36. However, the working group has previously refuted PCS as a match to the physical function domain as it is derived from all 8 domains of health-related quality of life (11). This provisional endorsement of SF-36 PF rather than PCS should prompt both researchers and the industry in reporting results for SF-36 PF in RCTs. Other areas needing more data for SF-36 PF are test-retest reliability and longitudinal construct validity.

Domain match for both instruments was rated as Amber. The inadequacy of domain match is perhaps related to the fact that these are generic instruments rather than instruments developed specifically for PsA. An issue of feasibility for the SF-36 PF was raised, as the entire SF-36 questionnaire must be administered for viable interpretation. SF-36 has been noted to be long and not being user friendly (11). The floor effect for both instruments has also been recognized, more prominently for HAQ-DI, and for those patients who have low disease activity (32). This limitation will affect the responsiveness of the instruments in PsA patients with stable disease. The instruments may be more useful in measuring response to treatment in patients with active disease and determining a worsening in patients with stable disease (20).

There are a few points worth mentioning in the interpretation of current data. Each instrument evaluation was performed separately using the OMERACT Filter 2.1, the fulfillment of which indicates that there is adequate evidence to support the instrument in clinical research. We made no comparison to demonstrate superiority of one instrument over the other. It is a possible that several instruments may fulfill the OMERACT Filter 2.1. Further work is needed to address which if any may be more useful in different research

settings; this is particularly true for RCT discrimination. Practically, RCTs have primarily reported HAQ-DI. When more data for SF-36 PF are available, the final rating may be upgraded. Comparison of responsiveness of instruments is certainly an unmet need that could be possibly addressed using a network meta-analysis method the working group is currently planning.

5. Conclusion

The body of evidence from this project demonstrated that both HAQ-DI and SF-36 PF fulfilled the necessary criteria for provisional endorsement as instruments to measure the physical function domain in PsA clinical trials. The instruments have been provisionally endorsed by the OMERACT TAG and GRAPPA members.

6. References

1. Ritchlin CT, Colbert RA, Gladman DD. Psoriatic Arthritis. N Engl J Med. 2017;376(21):2095-6.
2. Dures E, Hewlett S, Lord J, Bowen C, McHugh N, Group PS, et al. Important Treatment Outcomes for Patients with Psoriatic Arthritis: A Multisite Qualitative Study. Patient. 2017;10(4):455-62.
3. Orbai AM, de Wit M, Mease P, Shea JA, Gossec L, Leung YY, et al. International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. Annals of the rheumatic diseases. 2017;76(4):673-80.
4. Tillett W, Orbai AM, Ogdie A, Leung YY, Strand V, Gladman DD, et al. GRAPPA-OMERACT initiative to standardise outcomes in psoriatic arthritis clinical trials and longitudinal observational studies. Annals of the rheumatic diseases. 2018;77(5):e23.
5. Leung YY, Orbai AM, Ogdie A, Coates LC, de Wit M, Callis Duffin K, et al. The GRAPPA-OMERACT Psoriatic Arthritis Working Group at the 2018 Annual Meeting: Report and Plan for Completing the Core Outcome Measurement Set. J Rheumatol Suppl. 2019;95:33-7.
6. Leung YY, Tillett W, Orbai AM, Ogdie A, Eder L, Coates LC, et al. The GRAPPA-OMERACT Working Group: 4 Prioritized Domains for Completing the Core Outcome Measurement Set for Psoriatic Arthritis 2019 Updates. J Rheumatol Suppl. 2020;96:46-9.
7. 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.
8. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham III CO, Conaghan PG, et al. The OMERACT Handbook. 2017.
9. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.
10. Orbai AM, de Wit M, Mease PJ, Callis Duffin K, Elmamoun M, Tillett W, et al. Updating the Psoriatic Arthritis (PsA) Core Domain Set: A Report from the PsA Workshop at OMERACT 2016. J Rheumatol. 2017;44(10):1522-8.
11. Leung YY, Orbai AM, Ogdie A, Hojgaard P, Holland R, Goel N, et al. Appraisal of candidate instruments for assessment of the physical function domain in patients with psoriatic arthritis. J Rheumatol. 2020.
12. Hojgaard P, Klokker L, Orbai AM, Holmsted K, Bartels EM, Leung YY, et al. A systematic review of measurement properties of patient reported outcome measures in psoriatic arthritis: A GRAPPA-OMERACT initiative. Semin Arthritis Rheum. 2018;47(5):654-65.
13. Leung YY, Holland R, Mathew AJ, Lindsay C, Goel N, Ogdie A, et al. Clinical trial discrimination of physical function instruments for psoriatic arthritis: A systematic review. Semin Arthritis Rheum. 2020;50(5):1158-81.
14. Blackmore MG, Gladman DD, Husted J, Long JA, Farewell VT. Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modification. J Rheumatol. 1995;22(5):886-93.
15. Taccari E, Spadaro A, Rinaldi T, Riccieri V, Sensi F. Comparison of the Health Assessment Questionnaire and Arthritis Impact Measurement Scale in patients with psoriatic arthritis. Rev Rhum Engl Ed. 1998;65(12):751-8.
16. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum. 2007;57(5):723-9.

17. Leung YY, Tam LS, Kun EW, Ho KW, Li EK. Comparison of 4 functional indexes in psoriatic arthritis with axial or peripheral disease subgroups using Rasch analyses. J Rheumatol. 2008;35(8):1613-21.
18. Brodszky V, Pentek M, Balint PV, Geher P, Hajdu O, Hodinka L, et al. Comparison of the Psoriatic Arthritis Quality of Life (PsAQoL) questionnaire, the functional status (HAQ) and utility (EQ-5D) measures in psoriatic arthritis: results from a cross-sectional survey. Scand J Rheumatol. 2010;39(4):303-9.
19. Katchamart W, Benjamanukul S, Chiowchanwesawakit P. Validation of the Thai version of the Health Assessment Questionnaire for patients with psoriatic arthritis. Int J Rheum Dis. 2014;17(2):181-5.
20. Leung YY, Orbai AM, de Wit M, Balanescu A, Dernis E, Soubrier M, et al. Comparing the patient reported physical function outcome measures in a real-life international cohort of patients with psoriatic arthritis. Arthritis Care Res (Hoboken). 2020.
21. Wan MT, Walsh JA, Craig ET, Husni ME, Scher JU, Reddy SM, et al. A comparison of physical function instruments in psoriatic arthritis: HAQ-DI vs MDHAQ vs PROMIS10 global physical health. Rheumatology (Oxford). 2020.
22. Husted JA, Gladman DD, Farewell VT, Long JA, Cook RJ. Validating the SF-36 health survey questionnaire in patients with psoriatic arthritis. J Rheumatol. 1997;24(3):511-7.
23. Leung YY, Ho KW, Zhu TY, Tam LS, Kun EW, Li EK. Testing scaling assumptions, reliability and validity of medical outcomes study short-form 36 health survey in psoriatic arthritis. Rheumatology (Oxford). 2010;49(8):1495-501.
24. Leung YY, Tillett W, Hojgaard P, Orbai AM, Holland R, Mathew AJ, et al. Test-retest reliability for HAQ-DI and SF-36 PF for the measurement of physical function in psoriatic arthritis. J Rheumatol. 2021.
25. Husted JA, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. J Rheumatol. 1998;25(11):2146-55.
26. Leung YY, Zhu TY, Tam LS, Kun EW, Li EK. Minimal important difference and responsiveness to change of the SF-36 in patients with psoriatic arthritis receiving tumor necrosis factor-alpha blockers. J Rheumatol. 2011;38(9):2077-9.
27. Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales. J Rheumatol. 2010;37(5):1024-8.
28. Mease PJ, Woolley JM, Bitman B, Wang BC, Globe DR, Singh A. Minimally important difference of Health Assessment Questionnaire in psoriatic arthritis: relating thresholds of improvement in functional ability to patient-rated importance and satisfaction. J Rheumatol. 2011;38(11):2461-5.
29. Duarte-Garcia A, Leung YY, Coates LC, Beaton D, Christensen R, Craig ET, et al. Endorsement of the 66/68 Joint Count for the Measurement of Musculoskeletal Disease Activity: OMERACT 2018 Psoriatic Arthritis Workshop Report. J Rheumatol. 2019;46(8):996-1005.
30. Orbai AM, Holland R, Leung YY, Tillett W, Goel N, Christensen R, et al. PsAID12 Provisionally Endorsed at OMERACT 2018 as Core Outcome Measure to Assess Psoriatic Arthritis-specific Health-related Quality of Life in Clinical Trials. J Rheumatol. 2019;46(8):990-5.
31. Strand V, Boers M, Idzerda L, Kirwan JR, Kvien TK, Tugwell PS, et al. It's good to feel better but it's better to feel good and even better to feel good as soon as possible for as long as possible. Response criteria and the importance of change at OMERACT 10. J Rheumatol. 2011;38(8):1720-7.

32. Mease P, Strand V, Gladman D. Functional impairment measurement in psoriatic arthritis: Importance and challenges. Semin Arthritis Rheum. 2018;48(3):436-48.

7. Acknowledgement

We would like to thank Lara Maxwell and Dorcas Beaton, members of the OMERACT TAG in valuable advices and critically reviewed the body of evidence in this project. We thank all GRAPPA PRPs who participated in various workstreams. We thank all the GRAPPA members who read through the evidence and participated in the Delphi exercises for endorsement.

8. Conflict of interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

YYL is funded by the Clinician Scientist award of the National Medical Research Council, Singapore (NMRC/CSA-INV/0022/2017). AMO is funded by the Jerome L. Greene Foundation Scholar Award, the Staurulakis Family Discovery Award, the Rheumatology Research Foundation, and the National Institutes of Health (NIH) through the Rheumatic Diseases Resource-based Core Center (P30-AR053503 Cores A and D, and P30-AR070254, Cores A and B). PH and RC (The Parker Institute, Bispebjerg and Frederiksberg Hospital) is supported by a core grant from the Oak Foundation (OCAY-18-774-OFIL). LCC is funded by a National Institute for Health Research Clinician Scientist award, and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). AO is funded by NIH/NIAMS R01 AR072363.

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the funding agencies.

9. Author contributions

Ying Ying Leung: Conceptualization, Data curation, Analysis, Validation, Writing - original draft. **Ana-Maria Orbai:** Conceptualization, Data curation, Analysis, Validation, Writing – review & editing. **Pil Hojgaard:** Data curation, Analysis, Validation, Writing – review & editing. **Richard Holland:** Data curation, Analysis, Validation, Writing – review & editing. **Ashish J Mathew:** Data curation, Analysis, Validation, Writing – review & editing. **Niti Goel:** Conceptualization, Data curation, Analysis, Validation, Writing – review & editing. **Jeffrey Chau:** Data curation, Analysis, Validation, Writing – review & editing. **William Tillett:** Conceptualization, Data curation, Analysis, Validation, Writing – review & editing. **Chris Lindsay:** Data curation, Analysis, Writing – review & editing. **Alexis Ogdie:** Conceptualization, Writing – review & editing. **Laura C Coates:** Conceptualization, Writing - review & editing. **Dafna D Gladman:** Conceptualization, Writing - review & editing. **Robin Christensen:** Conceptualization, Validation, Writing - review & editing. **Philip Mease:** Conceptualization, Writing - review & editing. **Vibeke Strand:** Conceptualization, Writing - review & editing.

Table 1. Results of Delphi exercises with the working group and patient research partners for domain match and feasibility for HAQ-DI and SF-36 PF for PsA

		HAQ-DI			**SF-36 PF**		
	Response rate	% votes			% votes		
		Green	Amber	Red	Green	Amber	Red
Domain Match							
Working group votes:	13/13	46.2	**53.9**	0	38.5	**46.2**	15.4
PRP votes:	7/10	42.9	**42.9**	14.3	14.3	**57.1**	28.6
Final working group decision:	13/13	Amber			Amber		
Feasibility							
Working group votes:	13/13	**69.2**	30.8	0	30.8	**61.5**	7.7
PRP votes:	7/10	**71.4**	28.6	0	**57.1**	28.6	14.3
Final working group decision:	13/13	Green			Amber		

HAQ-DI: Health assessment questionnaire – disability index; PsA: psoriatic arthritis; PRP: patient research partners; SF-36 PF: physical functioning domain of Medical Outcome Survey Short Form -36 items.

Table 2. Summary of Measurement Property (SOMP) Table for HAQ-DI in PsA

Instrument: HAQ-DI				**Date completed: 24 Aug 2020**			
Author/year	**Truth** **Domain match**	**Feas-ibility**	**Truth**	**Discrimination**			
			Construct validity	Test- retest reliability	Long'l construct validity	Clinical trial discrimination	Thresholds of meaning
Working Group Appraisal (n=15)	+	+					
PRP Appraisal (n=7)	+	+					
Blackmore 1995			+/-				
Husted 1998					+		
Taylor 2007			+				
Kwok 2010							**+/-**
Mease 2011							+
Taccari 1998			+/-				
Leung 2008			+				
Brodszky 2010			+/-				
Leung 2011					+/-		+/-
Katchamart 2014			+				
New data from Leung 2016				+			
New data from Tillett				+			
New data from Leung 2020			+		+		+
New data from Wan 2020			+		+		
Data from 29 articles in 29 RCTs (n=29)						28 (+) 1 (+/-)	
Total available articles for each property			8	2	4	31	4
Total articles available for evidence synthesis			8	2	4	29*	4
Final Synthesis Rating	**AMBER From working group**	**GREEN From Working group**	**GREEN**	**GREEN**	**GREEN**	**GREEN**	**AMBER**
OMERACT Endorsement~~Overall rating for instrument across properties~~	**Provisionally endorsed: need additional work on domain match and threshold of meaning**						

Color coding for quality assessment: Green: Yes, likely low risk of bias; Amber: Some cautions but can be used as evidence; Red: No, don't use this evidence; White: no data.
(+): adequate performance; (+/-): equivocal; and (-): inadequate performance standards; *two studies rated as high risk of bias were not included in the evidence synthesis
Abbreviations. HAQ-DI: Health Assessment Questionnaire – Disability Index; PsA: psoriatic arthritis; PRP: patient research partner; SF-36 PF: physical Functioning subscale of Medical Outcome Survey Short-Form 36 items.

Table 3. Summary of Measurement Property (SOMP) Table for SF-36 PF subscale in PsA

Instrument: SF-36 PF				**Date completed: 24 Aug 2020**			
Author/year	**Truth** **Domain match***	**Feas-ibility***	**Truth** Construct validity	**Discrimination** Test-retest reliability	Long'l construct validity	Clinical trial discrimination	Thresholds of meaning
Working Group Appraisal (n=13)	+	+					
PRP Appraisal (n=7)	+	+					
Husted 1997			+				
Husted 1998					+		
Taylor 2007			+				
Leung 2008			+				
Leung 2010			+				
Leung 2011					+/-		+/-
New data from Tillett				+			
Kavanaugh 2006						**NA**	
Mease 2017						+/-	
Gladman 2017						+	
Mease 2018						**NA**	
Total available studies for each property	-	-	4	1	2	4	1
Total studies available for synthesis	-	-	4	1	2	2	1
Final Synthesis Rating ~~Rating (RAGW) [put on Master Checklist]~~	**Amber from working group**	**Amber from working group**	**GREEN**	**AMBER**	**AMBER**	**AMBER**	**AMBER**
OMERACT Endorsement~~Overall rating for instrument across properties~~	**Provisionally endorsed: need additional work on domain match, feasibility, test- retest reliability, longitudinal construct validity, clinical trial discrimination and threshold of meaning**						

Color coding for quality assessment: Green: Yes, likely low risk of bias; Amber: Some cautions but can be used as evidence; Red: No, don't use this evidence; White: no data.
(+): adequate performance; (+/-): equivocal; and (-): inadequate performance standards;
Abbreviations. HAQ-DI: Health Assessment Questionnaire – Disability Index; NA: not applicable; PsA: psoriatic arthritis; PRP: patient research partner; SF-36 PF: Physical Functioning subscale of Medical Outcome Survey Short-Form 36 items.