University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Measurement Properties of Radiographic Outcome Measures in Psoriatic Arthritis: A Systematic Review from the GRAPPA-OMERACT Initiative

Anna Antony[1], Richard Holland[2], Maria-Antonietta D'Agostino[3], Walter P. Maksymowych[4], Heidi Bertheussen[5], Lori Schick[6], Niti Goel[7], Alexis Ogdie[8], Ana-Maria Orbai[9], P. Højgaard[10], Laura C Coates[11], Vibeke Strand[12], Dafna D Gladman[13], Robin Christensen[14], Ying Ying Leung[15], Philip Mease[16] and William Tillett[17]

1. *School of Clinical Sciences, Monash University, Australia*
2. *Concord Repatriation Hospital, Australia*
3. *Versailles-Saint-Quentin University, Paris-Saclay INSERM U1173, Laboratoire d'Excellence INFLAMEX, France*
4. *University of Alberta, Canada*
5. *Patient Research Partner, Norway*
6. *Patient Research Partner, Canada*
7. *Patient Research Partner, Duke University School of Medicine, USA*
8. *University of Pennsylvania, USA*
9. *Johns Hopkins University School of Medicine, Division of Rheumatology, USA*
10. *Centre for Rheumatology and Spine Diseases, Rigshospitalet, Denmark*
11. *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, and Oxford Biomedical Research Centre, Oxford University Hospitals NHS Trust, Oxford, United Kingdom*
12. *Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto CA, USA*
13. *University of Toronto, Toronto Western Hospital, Canada*
14. *Musculoskeletal Statistics Unit, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen & Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark*
15. *Singapore General Hospital, Singapore*
16. *Swedish Medical Centre/Providence St. Joseph Health, USA*
17. *University of Bath, United Kingdom*

Corresponding Author: Dr. Anna Antony

Email Address: (anna.antony@monash.edu)

Postal Address:

Department of Rheumatology

Monash Medical Centre,

246 Clayton Road,

Clayton VIC 3168, Australia.

Abstract:

Background: Structural damage is as an important outcome in the Psoriatic Arthritis (PsA) Core Domain Set and its assessment is recommended at least once in the development of a new drug.

Objectives: To conduct a systematic review (SR) to identify studies addressing the measurement properties of radiographic outcome instruments for structural damage in PsA and appraise the evidence through the Outcome Measures in Rheumatology (OMERACT) Filter 2.1 Framework Instrument Selection Algorithm (OFISA).

Methods: A SR was conducted using search strategies in EMBASE and MEDLINE to identify full-text English studies which aimed to develop or assess the measurement properties of radiographic outcome instruments in PsA. Determination of eligibility and data extraction was performed independently by two reviewers with input from a third to achieve consensus. Two reviewers assessed the methodology and results of eligible studies and synthesized the evidence using OMERACT methodology.

Results: Twelve articles evaluating radiographic instruments were included. The articles assessed nine peripheral (hands, wrists and/or feet) and six axial (spinal and/or sacroiliac joints) radiographic instruments. The peripheral radiographic instruments with some evidence for reliability, cross-sectional construct validity and longitudinal construct validity were the Ratingen and modified Sharp van der Heijde scores. No instruments had evidence for clinical trial discrimination and thresholds of meaning. There was limited evidence for the measurement properties of all identified axial instruments.

Conclusion: There are significant knowledge gaps in the responsiveness of peripheral radiographic instruments. Axial radiographic instruments require further validation, and the need to generate novel instruments and utilise other imaging modalities should be considered.

Keywords: Psoriatic Arthritis, Imaging, Radiography, Outcome Instruments

INTRODUCTION

Structural damage in Psoriatic Arthritis (PsA) encompasses abnormalities in the structure or integrity of a joint, bone or tendon that may be attributable to PsA. Whilst there is significant heterogeneity in the phenotype of PsA patients, structural damage in randomised controlled trials (RCTs) has conventionally been measured using radiography of peripheral joints.

The 2016 Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA)-Outcome Measures in Rheumatology (OMERACT) Core Domain Set for PsA advocates that structural damage be measured at least once in the evaluation of a drug in randomised controlled trials (RCTs) and longitudinal observational studies (LOS). [1]

The OMERACT Filter 2.1 Instrument Selection Algorithm (OFISA) was developed in order to ensure that instruments used as outcome measures meet the three pillars of the OMERACT filter: truth, feasibility and discrimination. [2] Truth incorporates domain match, i.e., if the instrument has face, content, and construct validity. Feasibility considers factors such as cost, access, time taken to score, safety, knowledge transfer, and acceptability. Discrimination is determined by evaluating reliability and responsiveness (longitudinal construct validity, clinical trial discrimination and thresholds of meaning). [2] The steps in the OFISA include finding candidate instruments, assessing domain match and feasibility, and gathering and appraising the strength of the measurement properties for each instrument.

We conducted a systematic review (SR) of the published literature in order to determine candidate instruments for structural damage. We subsequently synthesized the current evidence for the measurement properties of available instruments and identified knowledge gaps to inform next steps for the research agenda.

METHODS

A protocol for a SR encompassing the measurement properties of outcome instruments for PsA was uploaded to PROSPERO (CRD42016032546). The GRAPPA-OMERACT working group has utilized modified versions of this protocol to conduct SRs to address other outcome domains such as patient-reported outcome measures. [3] The protocol has been adapted for use in this SR for radiographic outcome instruments, however the assessment of methods and measurement properties have been aligned with the novel OFISA methodology.

Literature search

A literature search limited to human studies was conducted by one reviewer (AA) in MEDLINE via PubMed from 1966 and EMBASE via OVID from 1974, both to 30 September 2019. The search strategy is included in the supplementary appendices (Table 1).

Eligibility criteria

Eligibility criteria were as follows: (1) The publication was a full-text original article in English, (2) The study sample represented the target population of either 100% PsA patients or ≥50% PsA patients if subgroup analyses were available, (3) The study aim was to develop or evaluate the measurement properties of a radiographic outcome instrument to assess structural damage, and (4) The radiographic outcome instrument was used to evaluate structural damage as an outcome. Studies that did not specifically aim to develop or evaluate measurement properties of a radiographic outcome instrument but that did report relevant data were considered indirect evidence and reported in the supplementary material only.

Selection of articles

The titles and abstracts were assessed by two independent reviewers (AA and WT). Full-text articles were reviewed where appropriate and article selection was by consensus. No additional studies were identified by co-authors. References were managed using Microsoft Excel.

Extraction of study characteristics and results

Two authors (AA and RH) independently extracted data regarding study design, population characteristics, and measurement properties. The results were summarized separately for peripheral instruments (hands, wrists and/or feet) and axial instruments (spinal and/or sacroiliac joints). The scoring proforma of instruments was summarized in the supplementary appendices (Tables 2 and 3).

Evaluation of the methodological quality per measurement property per study

Methodological quality was assessed using the COSMIN-OMERACT Good Methods Checklist (GMC). [4] Two reviewers (AA and WT) assessed the methodology independently and subsequently discussed discrepancies to achieve consensus. Studies were given a rating of 'Green' if good methods were used, 'Amber' if there were some methodological concerns but the data were acceptable for inclusion, and 'Red' if there was a high risk of bias.

Evaluating the adequacy of measurement properties per study

Each study was assessed using the OMERACT provisional standards (Supplementary Appendices Table 1) and assigned ratings of + (positive support for the measurement property), ± (ambivalent support, inconclusive), or – (instrument did not reach performance standards for that measurement property). [4] The syntheses of hypotheses required to generate ratings for construct validity and responsiveness were summarised (Supplementary appendices: Tables 7-10).

Synthesis of the evidence to ratings for the individual measurement properties of each radiographic outcome instrument

Studies with a high risk of bias (Red) were excluded from the final synthesis for each measurement property. The remaining studies were synthesized to generate an overall RED/AMBER/WHITE/GREEN (RAWG) rating for the individual measurement properties for each instrument based on the "Criteria for Final Rating" (Figure 1). [4] This rating summarises the quality and quantity of studies, the consistency of the results, and the performance of individual instruments. GREEN indicates 'Good to go', RED indicates 'Stop, do not continue', WHITE indicates 'No evidence' and AMBER indicates 'There is a concern, or caution, or weakness, but it is good enough to go forward perhaps with a research agenda to move it to GREEN or RED'. The results were summarized in a "Summary of Measurement Properties" table.

RESULTS

Study selection

The literature review yielded 9946 references (Figure 2). Of the 12 articles included, 7 evaluated peripheral instruments and 5 evaluated axial instruments. Articles with an adequate methodology (Green or Amber) that indirectly assessed the measurement properties of instruments (n=29) were summarized in supplementary tables (Tables 4 and 5).

PERIPHERAL RADIOGRAPHIC OUTCOME INSTRUMENTS

Study characteristics

The studies included were published between 1998 and 2019 (Table 1). The classification criteria used for PsA varied, however the study populations were sufficiently similar for studies to be considered together. Most studies were conducted in a single-centre and all but one were observational cohorts.

Characteristics of the radiographic outcome instruments

The peripheral radiographic outcome instruments were the Original Steinbrocker (oSteinbrocker) score, the Modified Steinbrocker (mSteinbrocker) score, the Modified Larsen (ML) score, the Ratingen score, two variations of the modified total Sharp scores (mTSS-A and –B), the modified Sharp van der Heijde (mSvdH) score, the Reductive X-Ray Score for Psoriatic Arthritis (ReXPsA) and the Simplified Psoriatic Arthritis Radiographic (SPAR) score (Supplementary Appendices Table 2).

The oSteinbrocker (score range 0-4), mSteinbrocker (0-168) and ML (0-250) instruments measure damage as a global score. The Ratingen score (0-360) measures destruction and proliferation separately. The mTSS-B (0-486) and mSvdH (0-528) instruments score the severity of erosions and joint space narrowing in the hands, wrists and feet, whilst the mTSS-A (0-386) only assesses joints in the hands and wrists. The ReXPsA (0-234) and SPAR (0-120) instruments assess proliferation, joint space narrowing and erosions, but use abbreviated scoring systems. [5, 6]

Feasibility

Feasibility data were infrequently reported other than scoring time, which was available for the mSteinbrocker, Ratingen, mTSS-B, mSvdH and SPAR scores (Table 2). [5, 7]

Inter- and Intra-Rater Reliability

Studies assessing cross-sectional inter- and intra-rater reliability were identified. The Ratingen instrument had ≥2 studies demonstrating good reliability (intra-class correlations of ≥0.70) and was rated AMBER. [5, 7-9] The OS, mSteinbrocker, ML, mTSS-A, mTSS-B, mSvdHs and SPAR instruments had good reliability in at least 1 study and were rated AMBER (Tables 2 and 4). [5, 7, 8, 10] Wassenberg et al. assessed the reliability of detecting change using the Ratingen score, but the ICC or Kappa was not calculated. [9]

Measurement Error

An instrument has an acceptable measurement error (+) if the smallest detectable change (SDC) or limits of agreement are less than the minimally important change (MIC). [11, 12] Measurement error has been assessed for the mSteinbrocker, mTSS-B, Ratingen and mSvdH instruments (Table 2). The MIC has not been defined for any instrument, therefore these instruments were rated AMBER for measurement error (Tables 2 and 4). [7]

Construct validity

Salaffi et al. evaluated the construct validity for the Ratingen, mSvdH and SPAR scores. [5] The Ratingen and mSvdH scores served as comparator instruments for the SPAR scores, and all instruments demonstrated good cross-sectional correlations as expected. Additional evidence for construct validity was available for the mSvdH score with a significant relationship demonstrated between radiographic damage and the Health Assessment Questionnaire-Disability Index (HAQ-DI) and Short Form-36 Physical Component Score (SF-36 PCS). [5] Ravindran et al. found that the mTSS-A had good correlations with 'clinical joint scores' and moderate correlations with the Health Assessment Questionnaire (HAQ). [10] The Ratingen, SPAR and mTSS-A scores received an AMBER rating while the mSvdHs was conferred a GREEN rating for this measurement property. (Tables 3 and 4). [10]

Longitudinal Construct Validity

The mSteinbrocker and Ratingen instruments had small effect sizes while the mTSS-B and mSvdH scores had moderate effect sizes when scoring was conducted in known chronology with a mean imaging interval of 25 months (Table 3). [7] Kerschbaumer et al. demonstrated heterogeneous results for the longitudinal construct validity of the mSvdHs (Table 3, Supplementary Appendices Table 10). [13] Rahman et al. used regression analyses to assess relative sensitivity to change for the OS, mSteinbrocker and ML scores. [8] The mSteinbrocker and ML scores were assigned a '±' rating while the OS score, which measures only a single affected joint or joint region, had a significantly lower sensitivity to change and was allocated a '-' rating. The ReXPsA instrument has only been assessed in the cohort from which this reductive score was derived, and the correlations demonstrated are therefore subject to confirmation bias. Following synthesis of the results and risk of bias, the OS was rated RED; the ML, mSteinbrocker, mTSS-B and mSvdHs were rated AMBER and the ReXPsA was rated WHITE (Table 4).

Clinical Trial Discrimination and Threshold of Meaning

All PsA RCTs have utilized either the mSvdHs or variants of the mTSS, reporting the differences in scores, differences in change scores, and/or proportion of patients with radiographic progression. [14-29] No studies meeting the requirements of the OFISA and SR protocol were identified. [4] Indirect evidence was summarized in the supplementary appendices (Tables 4 and 5).

AXIAL RADIOGRAPHIC OUTCOME INSTRUMENTS

Study characteristics

Five studies between 2007 and 2017 were included, encompassing six instruments. All studies were observational and used two differing definitions of 'axial PsA' ('AxPsA') between them.

Group A ('AxPsA'-A) met the ClASsification for PsA (CASPAR) and had inflammatory spinal pain (Calin criteria) and/or radiographic axial involvement (no formal definition provided). Group B ('AxPsA'-B) met CASPAR criteria, had ≥ unilateral grade 2 sacroiliitis (modified New York Criteria), and either inflammatory back pain or restricted spinal mobility (no definition provided for either). These studies have been synthesized separately given the potential differences in the underlying patient populations.

Characteristics of the radiographic outcome instruments

Instruments with reported measurement properties were the: modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS) score, Bath Ankylosing Spondylitis Radiology Index – Total (BASRI-T) score, Bath Ankylosing Spondylitis Radiology Index – Spine (BASRI-S) score, Psoriatic Arthritis Spondylitis Radiology Index (PASRI) score, mNYC and Radiographic Ankylosing Spondylitis Spinal Score (RASSS) scores. Of these, only PASRI and BASRI-T capture both the sacroiliac joints (SIJs) and vertebral spine within its scoring system, and only the PASRI assesses for involvement of the posterior elements. (Supplementary Appendices Table 3).

Feasibility

No formal estimation has been made regarding the time taken to score the individual instruments. Lubrano et al. reported that the mSASSS, BASRI-T and PASRI were feasible, while Biagioni et al. reported that all the components of the mSASSS, BASRI-S, PASRI, mNYC and RASSS instruments could be scored in a mean duration of 7 minutes by trained raters. [30-32]

Inter- and Intra-rater reliability

Reliability has not been assessed in the 'AxPsA'-A population. In 'AxPsA'-B, cross-sectional reliability was reported for the mSASSS, BASRI-Spine, PASRI, mNYC score and RASSS in one study. Intra-rater reliability was acceptable for all instruments, but inter-rater reliability

was only adequate for the PASRI (ICC >0.70). [32] All instruments were therefore rated AMBER for intra-rater reliability, and all instruments except the PASRI were rated RED for inter-rater reliability (Tables 2 and 5). No studies have assessed the reliability of image acquisition or reliability of detecting change in scores over time.

Measurement Error

Measurement error has not been reported for axial instruments.

Construct validity

In 'AxPsA'-A studies, good correlations were reported between the mSASSS, PASRI and BASRI-T scores, however correlations with patient-reported outcome measures and spinal metrology were moderate at best, which was an expected finding given these outcomes measure different constructs (Table 3). [30, 31] The mSASSS and PASRI had the strongest correlations with spinal metrology as measured by the Bath Ankylosing Spondylitis Metrology Index (BASMI). All three instruments were rated AMBER (Tables 3 and 5, Supplementary Appendices Table 7).

In 'AxPsA-B', moderate to good construct validity was demonstrated in a single study between the mSASSS score and spinal mobility, however the spinal mobility measurements used were a median of 10 assessments and the sample size was small (Table 3, Supplementary Appendices Table 7). [33] The mSASSS was allocated an AMBER rating in this population (Table 5).

Longitudinal Construct Validity

Longitudinal construct validity has been reported in one study in the 'AxPsA'-B population for the BASRI-S, mSASSS, RASSS and PASRI scores. In this study, a radiologist who was not blinded to chronology determined whether "true progression" occurred as a binary outcome. [34] The PASRI score increased in the greatest number of patients (32%), followed by the BASRI-S (29%) and mSASSS scores (25%); comparatively, "true progression" occurred in 24% of patients. The sensitivity and specificity for detecting "true progression" with a score increase of ≥1 with each instrument was comparable, sensitivity was highest with the mSASSS and PASRI, and the specificity was highest for the mSASSS and RASSS (Table 3). Overall, all four instruments had an acceptable specificity with a poor sensitivity when compared to a subjective comparator. These instruments' longitudinal construct validity were rated AMBER (Table 5).

Clinical Trial Discrimination and Threshold of Meaning

No published data on axial radiographic instruments were identified for patients with 'AxPsA'.

DISCUSSION

This systematic review summarises the measurement properties of peripheral and axial radiographic instruments, and informs the direction of future work necessary to select and endorse candidate instruments for the assessment of structural damage in PsA.

Assessing the peripheral instruments in turn, the Original Steinbrocker is inadequate given it only assesses the worst affected joint. Whilst its reliability is reasonable, its longitudinal construct validity is predictably poor. [8] The mSteinbrocker instrument assesses 42 joints in the hands, wrists and feet. A joint is scored 1 for the presence of juxta-articular osteopaenia, 2 if an erosion is present, 3 if there is co-existing joint space narrowing and erosion, and 4 if there is total joint destruction. The mSteinbrocker has been used to capture observational data in the Toronto PsA cohort, where it has been demonstrated to detect structural damage prior to it being clinically evident. [35] The simplified scoring translates to quicker scoring whilst maintaining good cross-sectional reliability in a single study. The main trade-offs are in its responsiveness and the inability to detect isolated joint space narrowing. [7] Similar limitations apply to the ML score, which additionally has no data for construct validity, and limited data for reliability and longitudinal construct validity. [8] The original Steinbrocker and ML instruments were developed from patients with severe destructive rheumatoid arthritis. Their use in the current era of early aggressive treatment is less relevant.

Proliferative changes are well-recognised in PsA, and include osteoproliferation (as captured in the Ratingen, ReXPsA and SPAR instruments), osteitis and ankyloses. The Ratingen instrument demonstrates cross-sectional reliability, has an acceptable measurement error and some evidence for its cross-sectional and longitudinal construct validity. [5, 7, 9] However, it does not assess joint space narrowing, which is an important albeit non-specific feature of PsA.

The ReXPsA and SPAR instruments assess erosions, joint space narrowing, and osteoproliferation individually. ReXPsA includes 22 joints and maintains the large scoring ranges of the mSvdH and Ratingen scores at the individual joints, but it has not been validated in a full-text publication outside of the cohort from which this instrument was derived. [6, 36] The SPAR instrument includes 40 joints, and each joint is assessed for the

presence of erosions, joint space narrowing and osteoproliferation as binary outcomes. While the instrument has demonstrated cross-sectional reliability and construct validity in a single study, its measurement error and longitudinal validity have not been assessed; the risk of a ceiling effect at an individual joint and potential lack of sensitivity to change are further concerns. [5]

Proliferative changes are important radiographic feature in classification of PsA, but the yield of measuring the progression of such features over time is uncertain. Tillett et al. reported a mean increase in osteoproliferation of 1.8 units/year using the Ratingen score in an observational cohort not stratified by treatment, suggesting that this is a feature that progresses over time. [7] Whilst there is some data on osteoproliferation in observational cohorts on biological therapy, no RCTs have directly utilised the Ratingen score. [37] A number of RCTs have assessed the yield of assessing for proliferative features such as osteitis and ankyloses, and have not noted a significant progression in these features over time nor a significant difference between treatment arms. [14, 15, 21, 22, 27, 38-41] These findings, and the impact on feasibility if such features were to be included, suggest that there may be little value in modifications of the mTSS or the mSvDHs to include proliferative change.

The mTSS-A, mTSS-B, and mSvdH instruments measure joint space narrowing and erosions. The utility of mTSS-A is limited as it only scores hand and wrist joints. In comparing the mTSS-B and mSvdH instruments, the latter has a larger scoring range, predominantly due to erosions being scored on either side of the joints in the feet. The mSvdH has also been more widely validated and is the only instrument other than the Ratingen score with an AMBER or GREEN rating in the domains of construct validity, cross-sectional inter- and intra-rater reliability, measurement error and longitudinal construct validity. Further strengths of the mSvdH instrument are its superior measurement error profile relative to the Ratingen and mSteinbrocker instruments, and the presence of post-hoc RCT data suggesting an association between the mSvdH and physical function as measured by the HAQ. [13]

The majority of placebo-controlled RCTs have demonstrated that radiographic progression as measured by the mSvdH and variants of the mTSS were significantly lower in intervention arms compared to placebo arms, with the exceptions reflecting the methodology of imputing missing data, the timepoint chosen for radiographic evaluation, the responsiveness of the radiographic instrument, or indeed a lack of difference in drug efficacy. [14-29] Furthermore, it is important to note that the mSvdHs has been successfully used to assess construct validity of composite outcome measures (supplementary material). However

'clinical trial discrimination' as per the current OFISA process, necessitates an a priori demonstration of an effect size between treatment arms or in a responder analysis. No studies in the available literature have reported effect sizes, and calculation of effect sizes based on published data may be problematic given radiographic progression data are likely to be non-parametric.

Similarly, no data exist for thresholds of meaning, which reflects the difficulties in defining what constitutes a MIC in an imaging instrument. [13] The MIC is the smallest change that is felt to be relevant by patients when assessed against an anchor such as pain or physical function. Kerschbaumer et al. in a post-hoc analysis of RCT data estimated that a patient in DAPSA (Disease Activity in Psoriatic Arthritis) remission with a mSvdHs of 150 would have a predicted residual mean HAQ (Health Assessment Questionnaire) score of 0.3, which is the MIC for HAQ in PsA. This is unfortunately not a useful corollary given a mSvdHs of 150 indicates severe radiographic damage uncommonly seen in RCTs, with the baseline mean SvdHs for PsA RCTs in the last decade ranging from 2.7-35.5. [14, 15, 21, 22, 27, 38-41] Given the slow rate of radiographic progression in RCTs, an MIC in its traditional definition may not be particularly relevant.

The synthesis of the peripheral instrument data favours the mSvdH score as a candidate instrument. The Ratingen, mSteinbrocker, SPAR and RexPsA instruments are potential alternatives, with their respective strengths and weaknesses as previously discussed. The main knowledge gaps that needs to be addressed moving forward are in the reliability of detecting change in peripheral instruments, which has only been demonstrated indirectly for the mSvdH score, defining the MIC, clinical trial discrimination, and thresholds of meaning. [41, 42] Consensus regarding domain match, face validity and feasibility within the OMERACT working group will aid in determining the final peripheral candidate instrument(s) selected for further evaluation to fill the identified knowledge gaps.

There is additional complexity in the synthesis of axial data, due to the use of non-standardised case definitions for 'AxPsA'; there is in fact no current consensus definition for 'AxPsA'. [43-47] Of the 5 axial instruments, only the BASRI-T and PASRI include the SIJs and vertebral spine, and only the PASRI was specifically developed for 'AxPsA'. While the extrapolation of radiographic instruments from the ankylosing spondylitis (AS) literature is practical given the significant overlap in radiographic features, there are some data to suggest variations in the symmetry and severity of sacroilitis, extent of lumbar involvement and morphology of syndesmophyte formation in 'AxPsA'. [45, 48]

Commented [AA1]: 2 Thoughts
1.Do you have access to an old RCT dataset? Is there value in specifically extracting out the HAQ questions on Grip + radiographic data from hands to look at this over time? Adjusting for disease activity in hands (maybe SJC only) of course. Tell me if this is a stupid idea. I'm sure there is a reason this has not been done.
2. Should we ask patients what is important to them as part of the domain match/feasibility exercise?

All assessed axial radiographic outcome instruments have been reported to be feasible in the literature, and this is supported by the routine collection of axial radiographic data in the Toronto PsA cohort. [30-32]

Reliability is an area of concern in axial instruments. In 'AxPsA-B', all instruments had acceptable cross-sectional intra-rater reliability, but inter-rater reliability was only acceptable for the PASRI (ICC = 0.88); all other instruments had ICCs between 0.52-0.68. [32] The same assessors found that these instruments seem to perform better in AS patients, potentially reflecting disease-specific factors, the older age of PsA patients leading to confounding due to osteoarthritis and diffuse idiopathic skeletal hyperostosis (DISH), and instrument-specific factors such as scoring of the posterior elements and a greater score range. [32] Whilst Lubrano et al. have published some data regarding "test-retest reliability", the scoring was performed by consensus among 3 assessors. There are no published data for the reliability of image acquisition or change in scores in 'AxPsA'.

Construct validity has been assessed for a number of axial instruments. Lubrano et al. has demonstrated good correlations between the BASRI-T, PASRI, and mSASSS, and predictably weak to moderate correlations between the radiographic instruments and the BASMI, Bath Ankylosing Spondylitis Functional Index (BASFI) and Revised Leeds Disability Questionnaire (RLDQ) in patients with 'AxPsA'-A. [30, 31] The mSASSS was also found to have moderate to excellent correlations with spinal metrology in a small group of patients with 'AxPsA'-B. [33]

Longitudinal construct validity has been investigated in one 'AxPsA'-B study, in which the BASRI-S, mSASSS, RASSS and PASRI were validated against assessment by an independent radiologist who was not blinded to chronology. The authors concluded that the PASRI appeared to have the best sensitivity in detecting radiographic progression, although all instruments performed poorly. The specificity of all instruments were good and comparable.

Determining candidate instrument(s) for 'AxPsA' based on current evidence is challenging. PASRI is the only instrument that has no RED ratings for its measurement properties, and only in the 'AxPsA'-B population. The axial instruments have no data for reliability of change scores, measurement error, clinical trial discrimination or thresholds of meaning in 'AxPsA'. The questionable cross-sectional reliability of these instruments in 'AxPsA' and the absence of longitudinal data is of concern given it is important that an instrument is able to reliably detect change within the time-frame of a clinical trial. The priorities moving forward will include the development of a standardised classification

criteria for axial involvement in Psoriatic Arthritis and considering if the use of novel instruments or alternate modalities are more appropriate prior to initiating further studies to address knowledge gaps.

This is the first systematic review of measurement properties of radiographic outcome measures in a PsA population, with the synthesis of evidence utilizing the OMERACT Filter 2.1 guidelines. [2] The standardized search strategy and grading of methodology and strength of evidence provided in these guidelines means that the process of updating literature searches will be streamlined in the future.

This review has a number of limitations. We have only identified instruments for assessing structural damage in the hands and wrists, feet, SIJs and spine. Contemporary evidence suggests that peripheral and axial radiographic damage is common and often progressive. [37, 46, 49, 50] However there is significant heterogeneity in the clinical phenotypes of PsA patients, particularly in observational cohorts. This raises the issue of monitoring structural damage in other phenotypes such as oligoarticular large joint PsA and enthesitis. The impact of large joint oligoarthritis is potentially captured to a degree in instruments that identify joint line tenderness in the absence of swelling, and functional impairment, but the lack of validated instruments to assess structural damage in oligoarthritis is a key unmet need. The original Steinbrocker could potentially be utilized in any peripheral joints, however it may be more appropriate for novel joint-specific imaging instruments to be developed. Entheseal structural damage would be more appropriately assessed via non-radiographic modalities.

Secondly, we have not made any recommendations in this paper on how instruments should be applied in RCTs. Important factors that warrant consideration include the number of readers involved, the blinding of readers to chronology and clinical information, whether serial radiographs should be scored in pairs, the score used (mean or through consensus), what outcome should be used (e.g. difference in mean change in score or proportion of patients who develop radiographic progression), what threshold of change should be considered as significant radiographic progression (e.g. any increase in score, an increase above the measurement error, or an increase above a yet un-defined minimally important difference) and the imputation of missing data. These strengths and limitations of these approaches in RCTs have been discussed elsewhere, and the impact of these approaches will be further assessed in context as part of our planned work on the clinical trial discrimination and thresholds of meaning. [28, 51] Standardisation of these approaches are important to ensure that results are comparable across clinical trials.

Inherent to the assessment of measurement properties of an instrument are that they should be ideally assessed in different subgroups to ensure validity within those groups. It is important to note that all the studies included in our literature review include patients with a mean disease duration of 2 years or more. Patients enrolled into PsA RCTs assessing the efficacy of biologics have typically had a mean or median disease duration exceeding 3 years, however it is likely that this will progressively shift to early disease in future trials. [14, 15, 21, 22, 27, 38-41] The value of radiographic endpoints in RCTs with predominantly early disease, particularly in comparison to magnetic resonance imaging (MRI) instruments, is an area that warrants additional research. [29, 52]

A further knowledge gap is the discriminative capacity of radiographic instruments to differentiate changes related to PsA from those related to primary osteoarthritis given the overlap in radiographic features such as joint space narrowing, the similar distribution of affected joints, and the potential for both diagnoses to co-exist. Indeed, these limitations extend to all instruments, including those that measure swollen and tender joints, and those that assess physical function and quality of life.

We have also limited our literature review to radiography. Parallel work streams are currently developing ultrasound and validating MRI instruments. [53-55] Whilst ultrasound and MRI may be more sensitive modalities for detecting structural progression, there are reciprocal issues related to sensitivity/specificity, clinical relevance, cost, access, standardization of image acquisition for centralized reading for RCTs, time taken to score and reliability.

The stratified synthesis of axial studies based on the different definitions of 'AxPsA' used does limit the number of studies available for synthesis, but this did not have a significant impact of final RAWG ratings for the individual measurement properties of each instrument (Supplementary Appendices Table 6). It is important to emphasize that a RAWG rating of RED or WHITE simply suggests that there is inadequate evidence or an absence of evidence at present to support the validity of the instrument.

Finally, we have excluded studies in which the a priori objective was not to specifically assess the measurement properties of radiographic outcome instruments. In all-inclusive analyses, these data provide important context and therefore have been included in the supplementary material.

CONCLUSION:

The measurement properties of instruments to assess structural damage in the peripheral joints have been reasonably validated, but a number of knowledge gaps need to be addressed in regard to domain match, feasibility and responsiveness. The measurement properties of axial radiographic outcome instruments require significant further validation and the need to generate novel instruments and/or utilise alternative imaging modalities should be considered.

References:

1. Orbai, A.M., et al., *Updating the Psoriatic Arthritis (PsA) Core Domain Set: A Report from the PsA Workshop at OMERACT 2016.* J Rheumatol, 2017. **44**(10): p. 1522-1528.
2. Beaton, D.E., et al., *Instrument Selection Using the OMERACT Filter 2.1: The OMERACT Methodology.* J Rheumatol, 2019. **46**(8): p. 1028-1035.
3. Hojgaard, P., 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): p. 654-665.
4. *Chapter 5: Instrument selection for Core Outcome Measurement Sets - Workbook.* OMERACT Handbook 2019 1st of March 2019; Available from: https://omeracthandbook.org/workbooks-%26-resources.
5. Salaffi, F., et al., *Preliminary validation of the Simplified Psoriatic Arthritis Radiographic Score (SPARS).* Skeletal Radiol, 2019. **48**(7): p. 1033-1041.
6. Tillett, W., et al., *Novel Composite Radiographic Score for Longitudinal Observational Studies of Psoriatic Arthritis: A Proof-of-concept Study.* J Rheumatol, 2016. **43**(2): p. 367-70.
7. Tillett, W., et al., *Feasibility, reliability, and sensitivity to change of four radiographic scoring methods in patients with psoriatic arthritis.* Arthritis Care Res (Hoboken), 2014. **66**(2): p. 311-7.
8. Rahman, P., et al., *Radiological assessment in psoriatic arthritis.* Br J Rheumatol, 1998. **37**(7): p. 760-5.
9. Wassenberg, S., et al., *A method to score radiographic change in psoriatic arthritis.* Z Rheumatol, 2001. **60**(3): p. 156-66.
10. Ravindran, J., et al., *A modified Sharp score demonstrates disease progression in established psoriatic arthritis.* Arthritis Care Res (Hoboken), 2010. **62**(1): p. 86-91.
11. Mokkink LB, D.A.M., *Protocol for performing a systematic review on imaging techniques.* 2017.
12. Prinsen, C.A.C., et al., *COSMIN guideline for systematic reviews of patient-reported outcome measures.* Qual Life Res, 2018. **27**(5): p. 1147-1157.
13. Kerschbaumer, A., et al., *The effects of structural damage on functional disability in psoriatic arthritis.* Ann Rheum Dis, 2017. **76**(12): p. 2038-2045.
14. Mease, P.J., et al., *Etanercept treatment of psoriatic arthritis: safety, efficacy, and effect on disease progression.* Arthritis Rheum, 2004. **50**(7): p. 2264-72.
15. Mease, P.J., et al., *Adalimumab for the treatment of patients with moderately to severely active psoriatic arthritis: results of a double-blind, randomized, placebo-controlled trial.* Arthritis Rheum, 2005. **52**(10): p. 3279-89.
16. Mease, P., et al., *Secukinumab improves active psoriatic arthritis symptoms and inhibits radiographic progression: primary results from the randomised, double-blind, phase III FUTURE 5 study.* Ann Rheum Dis, 2018. **77**(6): p. 890-897.
17. Mease, P.J., et al., *Secukinumab Inhibition of Interleukin-17A in Patients with Psoriatic Arthritis.* N Engl J Med, 2015. **373**(14): p. 1329-39.

18.   Mease, P.J., et al., *Etanercept and Methotrexate as Monotherapy or in Combination for Psoriatic Arthritis: Primary Results From a Randomized, Controlled Phase III Trial.* Arthritis Rheumatol, 2019. **71**(7): p. 1112-1124.

19.   Fraser, A.D., et al., *A randomised, double blind, placebo controlled, multicentre trial of combination therapy with methotrexate plus ciclosporin in patients with active psoriatic arthritis.* Ann Rheum Dis, 2005. **64**(6): p. 859-64.

20.   Kavanaugh, A., et al., *The Infliximab Multinational Psoriatic Arthritis Controlled Trial (IMPACT): results of radiographic analyses after 1 year.* Ann Rheum Dis, 2006. **65**(8): p. 1038-43.

21.   van der Heijde, D., et al., *Infliximab inhibits progression of radiographic damage in patients with active psoriatic arthritis through one year of treatment: Results from the induction and maintenance psoriatic arthritis clinical trial 2.* Arthritis Rheum, 2007. **56**(8): p. 2698-707.

22.   Kavanaugh, A., et al., *Golimumab in psoriatic arthritis: one-year clinical efficacy, radiographic, and safety results from a phase III, randomized, placebo-controlled trial.* Arthritis Rheum, 2012. **64**(8): p. 2504-17.

23.   Kavanaugh, A., et al., *Safety and Efficacy of Intravenous Golimumab in Patients With Active Psoriatic Arthritis: Results Through Week Twenty-Four of the GO-VIBRANT Study.* Arthritis Rheumatol, 2017. **69**(11): p. 2151-2161.

24.   Mease, P.J., et al., *Efficacy and safety of abatacept, a T-cell modulator, in a randomised, double-blind, placebo-controlled, phase III study in psoriatic arthritis.* Ann Rheum Dis, 2017. **76**(9): p. 1550-1558.

25.   Mease, P.J., et al., *Ixekizumab, an interleukin-17A specific monoclonal antibody, for the treatment of biologic-naive patients with active psoriatic arthritis: results from the 24-week randomised, double-blind, placebo-controlled and active (adalimumab)-controlled period of the phase III trial SPIRIT-P1.* Ann Rheum Dis, 2017. **76**(1): p. 79-87.

26.   Mease, P., et al., *Tofacitinib or Adalimumab versus Placebo for Psoriatic Arthritis.* N Engl J Med, 2017. **377**(16): p. 1537-1550.

27.   Kavanaugh, A., et al., *Ustekinumab, an anti-IL-12/23 p40 monoclonal antibody, inhibits radiographic progression in patients with active psoriatic arthritis: results of an integrated analysis of radiographic data from the phase 3, multicentre, randomised, double-blind, placebo-controlled PSUMMIT-1 and PSUMMIT-2 trials.* Ann Rheum Dis, 2014. **73**(6): p. 1000-6.

28.   van der Heijde, D., et al., *Effect of different imputation approaches on the evaluation of radiographic progression in patients with psoriatic arthritis: results of the RAPID-PsA 24-week phase III double-blind randomised placebo-controlled study of certolizumab pegol.* Ann Rheum Dis, 2014. **73**(1): p. 233-7.

29.   Coates, L.C., et al., *Effect of tight control of inflammation in early psoriatic arthritis (TICOPA): a UK multicentre, open-label, randomised controlled trial.* Lancet, 2015. **386**(10012): p. 2489-98.

30.   Lubrano, E., et al., *The radiological assessment of axial involvement in psoriatic arthritis: a validation study of the BASRI total and the modified SASSS scoring methods.* Clin Exp Rheumatol, 2009. **27**(6): p. 977-80.

31.   Lubrano, E., et al., *Psoriatic arthritis spondylitis radiology index: a modified index for radiologic assessment of axial involvement in psoriatic arthritis.* J Rheumatol, 2009. **36**(5): p. 1006-11.

32.   Biagioni, B.J., et al., *Reliability of radiographic scoring methods in axial psoriatic arthritis.* Arthritis Care Res (Hoboken), 2014. **66**(9): p. 1417-22.

33.   Chandran, V., et al., *Relationship between spinal mobility and radiographic damage in ankylosing spondylitis and psoriatic spondylitis: a comparative analysis.* J Rheumatol, 2007. **34**(12): p. 2463-5.

34.   Ibrahim, A., et al., *Sensitivity and Specificity of Radiographic Scoring Instruments for Detecting Change in Axial Psoriatic Arthritis.* Arthritis Care Res (Hoboken), 2017. **69**(11): p. 1700-1705.

35.   Siannis, F., et al., *Clinical and radiological damage in psoriatic arthritis.* Ann Rheum Dis, 2006. **65**(4): p. 478-81.

36.   Isnardi, C., et al., *Validation of the Rexspa (Reductive X-Ray Score for Psoriatic Arthritis) in an Argentinean Cohort of Patients with Psoriatic Arthritis.* Arthritis Rheumatol, 2018. **70**(suppl 10).

37.   Allard, A., et al., *Trajectory of radiographic change over a decade: the effect of transition from conventional synthetic disease-modifying antirheumatic drugs to anti-tumour necrosis factor in patients with psoriatic arthritis.* Rheumatology (Oxford), 2019. **58**(2): p. 269-273.

38.   Mease, P.J., et al., *Continued inhibition of radiographic progression in patients with psoriatic arthritis following 2 years of treatment with etanercept.* J Rheumatol, 2006. **33**(4): p. 712-21.

39.   Kavanaugh, A., et al., *Effect of infliximab therapy on employment, time lost from work, and productivity in patients with psoriatic arthritis.* J Rheumatol, 2006. **33**(11): p. 2254-9.

40.   Antoni, C.E., et al., *Two-year efficacy and safety of infliximab treatment in patients with active psoriatic arthritis: findings of the Infliximab Multinational Psoriatic Arthritis Controlled Trial (IMPACT).* J Rheumatol, 2008. **35**(5): p. 869-76.

41.   Kavanaugh, A., et al., *Radiographic Progression Inhibition with Intravenous Golimumab in Psoriatic Arthritis: Week 24 Results of a Phase III, Randomized, Double-blind, Placebo-controlled Trial.* J Rheumatol, 2019. **46**(6): p. 595-602.

42.   Kavanaugh, A., et al., *Clinical efficacy, radiographic and safety findings through 5 years of subcutaneous golimumab treatment in patients with active psoriatic arthritis: results from a long-term extension of a randomised, placebo-controlled trial (the GO-REVEAL study).* Ann Rheum Dis, 2014. **73**(9): p. 1689-94.

43.   Yap, K.S., et al., *Back pain in psoriatic arthritis: defining prevalence, characteristics and performance of inflammatory back pain criteria in psoriatic arthritis.* Ann Rheum Dis, 2018. **77**(11): p. 1573-1577.

44.   Fernandez-Sueiro, J.L., et al., *Validity of the bath ankylosing spondylitis disease activity index for the evaluation of disease activity in axial psoriatic arthritis.* Arthritis Care Res (Hoboken), 2010. **62**(1): p. 78-85.

45.   Jadon, D.R., et al., *Axial Disease in Psoriatic Arthritis study: defining the clinical and radiographic phenotype of psoriatic spondyloarthritis.* Ann Rheum Dis, 2017. **76**(4): p. 701-707.

46.   Queiro, R., et al., *Clinically asymptomatic axial disease in psoriatic spondyloarthropathy. A retrospective study.* Clin Rheumatol, 2002. **21**(1): p. 10-3.

47.   Feld, J., et al., *Axial disease in psoriatic arthritis and ankylosing spondylitis: a critical comparison.* Nat Rev Rheumatol, 2018. **14**(6): p. 363-371.

48.   Helliwell, P.S., P. Hickling, and V. Wright, *Do the radiological changes of classic ankylosing spondylitis differ from the changes found in the spondylitis associated with inflammatory bowel disease, psoriasis, and reactive arthritis?* Ann Rheum Dis, 1998. **57**(3): p. 135-40.

49.   Chandran, V., et al., *Axial psoriatic arthritis: update on a longterm prospective study.* J Rheumatol, 2009. **36**(12): p. 2744-50.

50.   Touma, Z., et al., *Clinical and Demographic Characteristics of Erosion-free and Erosion-present Status in Psoriatic Arthritis in a Cohort Study.* J Rheumatol, 2016. **43**(6): p. 1057-62.

51.   van der Heijde, D., et al., *Presentation and analysis of data on radiographic outcome in clinical trials: experience from the TEMPO study.* Arthritis Rheum, 2005. **52**(1): p. 49-60.

52.   Helliwell, P.S., et al., *Comparing Psoriatic Arthritis Low-field Magnetic Resonance Imaging, Ultrasound, and Clinical Outcomes: Data from the TICOPA Trial.* J Rheumatol, 2020. **47**(9): p. 1338-1343.

53.   Boyesen, P., et al., *The OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Score (PsAMRIS) is reliable and sensitive to change: results from an OMERACT workshop.* J Rheumatol, 2011. **38**(9): p. 2034-8.

54.   Glinatsi, D., et al., *Validation of the OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Score (PsAMRIS) for the Hand and Foot in a Randomized Placebo-controlled Trial.* J Rheumatol, 2015. **42**(12): p. 2473-9.

55.   Terslev, L., et al., *The OMERACT Stepwise Approach to Select and Develop Imaging Outcome Measurement Instruments: The Musculoskeletal Ultrasound Example.* J Rheumatol, 2019. **46**(10): p. 1394-1400.