Seminars in Arthritis and Rheumatism 51 (2021) 367–386

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

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

Measurement properties of radiographic outcome measures in Psoriatic Arthritis: A systematic review from the GRAPPA-OMERACT initiative

Anna Antony[a,*], Richard Holland[b], Maria-Antonietta D'Agostino[c], Walter P. Maksymowych[d], Heidi Bertheussen[e], Lori Schick[f], Niti Goel[g], Alexis Ogdie[h], Ana-Maria Orbai[i], P. Højgaard[j], Laura C Coates[k], Vibeke Strand[l], Dafna D. Gladman[m], Robin Christensen[n], Ying Ying Leung[o], Philip Mease[p], William Tillett[q]

[a] School of Clinical Sciences, Department of Rheumatology, Monash Medical Centre, Monash University, 246 Clayton Road, Clayton 3168 VIC, Australia
[b] Department of Rheumatology, Concord Repatriation General Hospital, New South Wales, Australia
[c] Paris-Saclay INSERM U1173, Laboratoire d'Excellence INFLAMEX, Versailles-Saint-Quentin University, France
[d] University of Alberta, Canada
[e] Patient Research Partner, Norway
[f] Patient Research Partner, Canada
[g] Patient Research Partner, Duke University School of Medicine, United States
[h] Division of Rheumatology, University of Pennsylvania, Pennsylvania, United States
[i] Division of Rheumatology, Johns Hopkins University School of Medicine, United States
[j] Centre for Rheumatology and Spine Diseases, Rigshospitalet, Denmark
[k] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, and Oxford Biomedical Research Centre, Oxford University Hospitals NHS Trust, Oxford, United Kingdom
[l] Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto CA, United States
[m] Krembil Research Institute, University of Toronto, Toronto Western Hospital, Canada
[n] 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
[o] Department of Rheumatology and Immunology, Singapore General Hospital, Singapore
[p] Rheumatology, Swedish Medical Center, Seattle, Washington, USA
[q] Department of Pharmacy and Pharmacology, University of Bath, Bath, United Kingdom

ARTICLE INFO

Keywords:
Psoriatic Arthritis
Imaging
Radiography
Outcome instruments

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 or 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.

* Corresponding author.
E-mail address: anna.antony@monash.edu (A. Antony).

https://doi.org/10.1016/j.semarthrit.2021.01.008
0049-0172/

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" (Fig. 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 (Fig. 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).

Table 1
A and B: Study design, demographics and radiographic outcome instruments.

Peripheral Radiographic Studies

Study	Country	Study Design and Population Selection (Classification Criteria, if reported)	Sample Size and Intervention	AgeMean (SD) Median [IQR]	Sex	Disease Duration (years)Mean (SD)Median [IQR]	Radiographic SeverityMean (SD) Median [IQR]	Disease ActivityMean (SD) Median [IQR]	ESR or CRPMean (SD) Median [IQR]	Blinding			Radiographic DurationMean (SD)Median [IQR]	Radiographic Outcome Instrument	Joints Read
										Clinical Data	Chronology	Paired Analysis			
Rahman 1998	Canada	Retrospective cohort Radiographs selected to represent a spectrum of radiographic damage	68 patients	40.86 (12.42) at presentation to clinic	NR	6.19 (8.70)	NR	Number of active joints 9.36 (7.73) Number of effusions 3.15 (3.03)	NR	YES	YES	NO	≥2 Years	MS Score OS Score ML Score	Hands, Wrists and Feet
Wassenberg 2001	Germany	Retrospective cohort Consecutive patients	20 patients commencing Methotrexate	47.8 [median]			14/20 M	5.2 [0.5–23]	NR	NR	NR	YES	NO	YES	3 years
Ratingen Score	Hands, Wrists and Feet														
Ravindran 2010	United Kingdom	Retrospective cohort Convenience sample (Moll and Wright)	139 patients	45 (13.4)	66/139 M	5 [2.0–15.0]	mTSS-A 4.0 [0.0–31.0] ERO 1.0 [0.0–10] JSN 2.0 [0.0–15.0]	Clinical joint score (swelling or deformity in 70 joints) 6.5 [2.2–17.0]	NR	YES	YES	NR	5.75 years [Median]	mTSS-A	Hands and Wrists
Tillett 2014	United Kingdom	Retrospective cohort Consecutive patients (CASPAR Criteria)	50 patients commencing a TNF-inhibitor	50 (12.1)	NR	10 (8.4)	MS Score 15.4 (21.63) Ratingen Score 13.2 (25.23) mTSS-B 26.3 (39.05) mSvdHS 26.8 (38.25)	NR	NR	YES	NO	YES	25 (9.6) months	MS Score Ratingen Score mTSS-B mSvdHS	Hands, Wrists and Feet
Tillett 2016	United Kingdom	Retrospective cohort Consecutive patients (CASPAR Criteria)	50 patients commencing a TNF-inhibitor	50 (12.1)	NR	10 (8.4)	MS Score 15.4 (21.63) Ratingen Score 13.2 (25.23) mTSS-B 26.3 (39.05) mSvdHS 26.8 (38.25)	NR	NR	YES	NO	YES	2.1 years (mean)	ReXPsA mTSS-B Ratingen Score mSvdHS	Hands, Wrists and Feet
Kerschbaumer 2017 **GO-REVEAL**	6 countries	Randomised Controlled Trial – Post-Hoc Analysis (CASPAR criteria) Remission (DAPSA ≤4) Major response (85% DAPSA improvement from baseline and HAQ baseline ≥1)	363 patients in total 117 patients 67 patients	46.9 (10.8) 44.0 (11.5) 43.6 (11.2)	210/363 M 78/117 M 37/67 M	7.4 (7.4) 7.2 (6.7) 7.8 (8.3)	mSvdHs Total: 9.5 [3–26] 9.5 [3–26] 12 [4–56.2]	66/68 Joint Count 13.3 (10.3)/23.1 (16.5) 11.1 (8.3)/16.7 (11.4) 17.1 (11.7)/29.3 (17.6) DAPSA/c-DAPSA also reported	CRP (mg/dL) 1.4 (1.6) 1.4 (1.6) 2.2 (1.9)	YES	YES	NO	104 weeks	mSvdHS	Hands, Wrists and Feet
	Austria	Retrospective cohort Convenience sample	160 patients 55 in remission	52.3 (12) 51.8 (12.1)	85/160 M 40/55 M	2.9 (7.1) 3.6 (8.7)	6 [2–14] 8 [2–21]	66/68 Joint Count 2.7 (3.5)/ 10.1 (14.5) 2.3 (3.4)/ 4.2 (8.9) c-DAPSA also reported	CRP (mg/dL) 0.9 (0.8) 0.5 (0.3)	NR	NR	NR	NR	mSvdHS	Hands, Wrists and Feet
Salaffi 2019	Italy	Cross-sectional Consecutive patients with peripheral joint involvement (CASPAR Criteria)	105 patients	50.2 (12.1)	34/105 M	10.1 (8.4)	SPARS: 51.55 [43.48–57.00] mSvdHS:245.00 [217.90–275.01] PARS: 156 [133.90–167.01] Means reported	66/68 Joint Count 6 [Range: 0–11]/ 8 [Range: 0–31]	CRP (mg/dL) 0.7 [Range: 0.1–8.7]	NR	N/A	N/A	NR	SPARS mSvdHs Ratingen Score	Hands, Wrists and Feet

Axial Radiographic Studies

Study	Country	Study Design and Population Selection (Classification Criteria, if reported)	Sample Size and Intervention	AgeMean (SD) Median [IQR]	Sex	Disease Duration (years) Mean (SD) Median [IQR]	Radiographic Severity Mean (SD) Median [IQR] or *[range]*	Spinal Measurements Mean (SD) Median [IQR] or *[range]*	Blinding			Radiographic Duration Mean (SD) Median [IQR]	Radiographic Outcome Instrument	Joints Read
									Clinical Data	Chronology	Paired Analysis			
Chandran 2007	Canada	Cross-sectional Convenience sample (PsA and ≥ Grade 2 sacroiliitis)	10 patients with AxPsA 9 patients with Ankylosing Spondylitis	52 (Mean) 38 (Mean)	9/10 M 7/9 M	17 (mean) 16 (mean)	mSASSS/ BASRI-S 13 [0–46]/ 7 [1–10.5] 16.6 [1–72]/ 7 [5–12]	Occiput-to-wall 5.5 [0–17] cm Tragus-to-wall 15.6 [11.6–26.5] cm Cervical rotation 56 [7.5–82] degrees Chest expansion 3.6 [1.5–7.7] cm Modified Schober's 4.3 [1.1–6.4] cm Domjan spinal flexion 14.6 [5.3–22.4] cm INSPIRE spinal flexion 30.1 [9.9–42.3] cm Occiput-to-wall 7 [0–16.5] cm Tragus-to-wall 16.8 [10.9–26.5] cm Cervical rotation 49 [33.8–76] degrees Chest expansion 3.1 [1.8–4] cm Modified Schober's 3 [0.5–5.9] cm Domjan spinal flexion 11.8 [3.5–17.3] cm INSPIRE spinal flexion 25.3 [7.5–35] cm	YES	N/A	N/A	N/A	mSASSS BASRI-S	AP Pelvis AP and Lateral Cervical and Lumbar Spine
Lubrano 2009 PASRI	Italy	Cross-sectional Consecutive patients (CASPAR Criteria + Spinal inflammatory pain (Calin) AND/OR 'Radiologic Axial Involvement')	73 patients	49.4 (11.0)	54/73 M	14.0 (7.9)	BASRI-T 2.25 [0–14] mSASSS 0 [0–42] PASRI 4 [0–64]	Occiput-to-wall 2 [0–24] cm Tragus-to-wall 13 [7–28] cm Cervical rotation 45 [0–90] degrees Chest expansion 3.3 [0.5–5.5] cm Modified Schober's 4 [0–9] cm Intermalleolar distance 98 [45–126] cm Finger to floor 20 [0–70] cm BASMI 3 [0–8]	NR	YES	N/A	N/A	mSASSS BASRI-T PASRI	AP Pelvis AP and Lateral cervical, Dorsal and Lumbar Spine
Lubrano 2009	Italy	Cross-sectional Consecutive patients (CASPAR Criteria + Spinal inflammatory pain (Calin) AND/OR 'Radiologic Axial Involvement')	77 patients	49.4 (10.8)	58/77 M	13.9 (7.9)	Not reported	BASMI 3 [0–8]	YES	N/A	N/A	N/A	mSASSS BASRI-T	AP Pelvis AP and Lateral Thoracic and Lumbar Spine Lateral Cervical Spine
Biagioni 2014	Canada	Cross-sectional Convenience sample (CASPAR Criteria + ≥ Unilateral Grade 2 Sacroilitis + Inflammatory Back Pain OR Restricted Spinal Mobility)	40 patients with Axial PsA 18 patients with Ankylosing Spondylitis	53 (14) 45 (12)	24/40 M 12/18 M	18 (9.7) 12 (12.1)	BASRI-S: 3.98 (2.38) mSASSS: 8 (13.4) RASSS: 6.54 (14.1) PASRI: 12 (12.3) BASRI-S: 4.83 (3.13) mSASSS: 11 (18.5) RASSS: 8.26 (16.32) PASRI: 18.3 (17.7)	Occiput-to-wall 2.3 (4.3) cm Cervical rotation 61 (21) degrees Chest expansion 5.6 (1.9) cm Domjan spinal flexion 16 (5.2) cm Modified Schober's 4.7 (2.7) cm Intermalleolar distance 104 (20.5) cm Occiput-to-wall 7.4 (7.4) cm Cervical rotation 37 (21) degrees Chest expansion 4.3 (2.3) cm Domjan spinal flexion 10.5 (5.9) cm Modified Schober's 4.9 (5.3) cm Intermalleolar distance 100 (16.4) cm	YES	N/A	N/A	N/A	mNYC mSASSS BASRI-S PASRI RASSS	AP Pelvis AP and Lateral Thoracic and Lumbar Spine Lateral Cervical Spine
Ibrahim 2017	Canada	Retrospective cohort Convenience sample CASPAR Criteria + ≥ Unilateral Grade 2 Sacroilitis + Inflammatory Back Pain OR Restricted Spinal Mobility	105 patients	51.9 (13.7)	71/105 M	16.0 (10)	BASRI-S: 3.4 (2.1) mSASSS: 3.4 (8.0) RASSS: 3.6 (9.4) PASRI: 8.5 (9.3)	Occiput-to-wall 1.9 (4.2) cm Chest Expansion 6 (2.1) cm Lumbar lateral flexion 16.3 (4.9) cm Schober's test 4.5 (1.4) cm Intermalleolar distance 99.3 (26.4) cm	YES	YES	NO	≥2 years	mSASSS BASRI-S PASRI RASSS	AP Pelvis AP and Lateral Thoracic and Lumbar Spine Lateral Cervical Spine

(MS: Modified Steinbrocker; OS: Original Steinbrocker; ML: Modified Larsen; mTSS-A: modified Total Sharp Score-A; mTSS-B: modified Total Sharp Score-B; mSvdHs: modified Sharp van der Heidje score; ReXPSA: Reductive X-Ray Score for Psoriatic Arthritis; SPARS: Simplified Psoriatic Arthritis Radiographic Score; mSASSS: modified Stoke Ankylosing Spondylitis Spinal Score; BASRI-S: Bath Ankylosing Spondylitis Radiology Index – Spine; BASRI-T: Bath Ankylosing Spondylitis Radiology Index – Total; PASRI: Psoriatic Arthritis Spondylitis Radiology Index; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; mNYC: modified New York Criteria; c-DAPSA: Clinical Disease Activity Index for Psoriatic Arthritis; DAPSA: Disease Activity Index for Psoriatic Arthritis; CASPAR: Classification for Psoriatic Arthritis; HAQ: Health Assessment Questionnaire; TNF: Tumour Necrosis Factor; AxPsA: Axial Psoriatic Arthritis; NR: Not reported; N/A: Not applicable; CRP: C-Reactive Protein; AP: Anterior-Posterior).

Tables 2
A and B: Reliability and feasibility.

(A) Peripheral Radiographic Instruments

Study	ROI	No. of radiographs	Interval for Reliability	Observers stable? (n)	Test conditions stable?	Intra-Observer Reliability ICC (95% CI)	Inter-Observer ReliabilityICC (95% CI)	Measurement error	Feasibility	Judgement and notes
Rahman 1998	OS Score MS Score ML Score	68 for inter-rater 20 for intra-rater	Not reported	Yes (2)	NR	ICC: Two-way, mixed effects model, agreement, multiple raters 0.90 (0.74–0.96), 0.86 (0.65–0.95) 0.80 (0.52–0.92), 0.81 (0.59–0.93) 0.84 (0.62–0.94), 0.85 (0.64–0.95)	ICC: Two-way, mixed effects model, agreement, single rater 0.86 (0.76–0.90) 0.86 (0.76–0.90) 0.87 (0.79–0.92)		'Easy to score, not time consuming and relatively inexpensive'	**Intra-observer: OS, MS and ML +** **Inter-observer: OS, MS and ML +** Radiographs selected to represent a spectrum of radiographic damage
Wassenburg 2001	Ratingen Score	40	≥4 weeks	Yes (2)	Yes	ICC for status scores not calculated. Graph used to demonstrate agreement and subjective decision made regarding reliability. **Total score:** Rater 1: 'Very good agreement' Rater 2: 'Good agreement' **Destruction score:** Rater 1: 'Good agreement' Rater 2: 'Good agreement' **Proliferation score:** Rater 1: 'Acceptable agreement' Rater 2: 'Good agreement' **Hierarchical analysis of variance model used to assess reliability of change scores** **Variance** Destruction: 2.8 and 5.1 Proliferation: 3.5 and 2.3 Total Score: 4.5 and 7.0 **Reliability** Destruction: 3.3 and 2.0 Proliferation: 2.2 and 4.2 Total Score: 3.6 and 2.8	ICC for status scores not calculated. Graph used to demonstrate agreement and subjective decision made regarding reliability. **Total Score:** Reading 1: 'Very good' Reading 2 'Good agreement' **Hierarchical analysis of variance model used to assess reliability of change scores** **Variance:** Destruction: 4.0 Proliferation: 2.9 Total Score: 5.6 **Reliability:** Destruction score: 3.9 Proliferation score: 2.8 Total score: 4.1	**MDC** **Rater 1 / Rater 2 / Intra-rater** **Destruction** 7.8 / 14.3 / 11.5 **Proliferation** 8.9 / 6.4 / 8.4 **Total** 12.6 / 19.6 / 16.5 **MDC% (of total score/sub-score)** **Rater 1 / Rater 2 / Intra-rater** **Destruction** 3.9 / 7.2 / 5.8 **Proliferation** 6 / 4 / 5 **Total** 3.5 / 5.4 / 4.6		**Intra-observer (status scores and change scores): ±** **Inter-observer (status scores and change scores): ±** **Measurement error: ± (MIC not defined)** A Hierarchical analysis of variance model was used where 3 variability components were estimated using mean-square errors: - Variance over time (reflecting radiographic progression) - Inter-rater variance - Intra-rater variance Intra-rater reliability = √Radiographic change SD divided by √Intra-rater SDInter-rater reliability = √Radiographic change SD divided by √Inter-rater SD (Higher values indicated higher reliability and a ratio of 1 indicated that all change was attributable to measurement error)
Ravindran 2010	mTSS-A	10	1 month	Yes (2 for Inter- and 1 for Intra-rater reliability)	NR	ICC model/type N/A 0.99 (0.99–0.99)	ICC model/type Method N/A 0.99 (0.98–0.99)		Training was "time-consuming"	**Intra-observer: +** **Inter-observer: +** Reliability scoring undertaken by readers not involved in scoring of the radiographs for construct validity of the instrument
Tillett 2014	MS Score	10	1 month	Yes (2)	Yes	ICC: Two-way, mixed effects model,	ICC: Two-way, mixed effects model,	**SDC (%)** **SDD (%)** **SEM**	**Mean time: (minutes)**	**Intra-observer: MS +, mTSS-*B* +, mSvdHs +, Ratingen +**

(*continued on next page*)

Tables 2 (*Continued*)

(A) Peripheral Radiographic Instruments

Study	ROI	No. of radiographs	Interval for Reliability	Observers stable? (n)	Test conditions stable?	Intra-Observer Reliability ICC (95% CI)	Inter-Observer ReliabilityICC (95% CI)	Measurement error			Feasibility	Judgement and notes
						agreement, single raters	agreement, multiple raters	4.83 (2.87)	8.11 (4.82)	3.49	6.2	**Inter-observer: MS -, mTSS-*B* +, mSvdHs +, Ratingen +**
						0.99 (0.95–1.00), 1.00 (0.99–1.00)	0.42 (−0.21–0.81), 0.40 (−0.23–0.81) **					**Measurement error: MS ±, mTSS-*B* ±, mSvdHs ±, Ratingen ± (MICs not defined)**
	mTSS-B					0.99 (0.95–1.00), 1.00 (0.99–1.00)	0.94 (0.78–0.96), 0.96 (0.86- 0.99)	7.01 (1.44) 4.11	11.77 (2.42) 6.90	5.06 2.97	14.6	Inter-rater reliability assessment for MS Score was subsequently conducted in 50 patients – ICC improved to 0.88 (0.77, 0.84) due to less osteopaenia compared to original sample
	ERO					0.77 (0.35–0.94), 1.00 (0.99–1.00)	0.77 (0.35–0.94), 0.64 (0.10–0.90)	5.05	8.47	3.64		
	JSN					0.95 (0.81–0.99), 0.99 (0.98–1.00)	0.96 (0.86–0.99) 0.94 (0.81–0.99)					
	mSvdHs					0.97 (0.90–0.99), 0.99 (0.98–0.99)	0.95 (0.83–0.99), 0.99 (0.96 –1.00)	6.45 (1.22) 4.36	10.83 (2.05) 7.31	4.66 3.14	14.4	The SDD% and SDC% results favour the mTSS-B and mSvdHS, but the smallest detectable change was greater than the mean change over 2 years for all instruments.
	ERO					0.91 (0.69–0.98), 1.00 (0.99–1.00)	0.91 (0.70–0.98), 0.92 (0.72–0.98)	4.35	7.29	3.14		
	JSN					0.93 (0.76–0.98), 0.98 (0.97–0.99)	0.96 (0.87–0.99), 0.92 (0.73–0.98)					
	Ratingen Score					0.99 (0.95–1.00), 0.99 (0.97–1.00)	0.89 (0.64–0.97), 0.90 (0.65–0.97)	7.57 (2.10) 3.48	12.71 (3.53) 5.83	5.46 2.51	10.5	
	Destruction					0.76 (0.31–0.93), 1.00 (0.99–1.00)	0.69 (0.18–0.91), 0.69 (0.18–0.91)	4.37	7.34	3.15		
	Proliferation					0.90 (0.66–0.97), 1.00 (0.98–1.00)	0.90 (0.67–0.97), 0.85 (0.52–0.96)					
Salaffi 2019	SPARS	105	2 weeks	Yes (2)	Yes	ICC: One-way, random effects model, consistency, single raters	ICC: One-way, random effects model, consistency, multiple raters	SDD for the average of the SPARS = 8.0			4.5 [Range 3.2–6.9] mins	ICCs to be interpreted with caution given ICC model used may have over-estimated reliability
	mSvdHS					0.95 (0.92–0.97)	0.884 (0.852–0.898)				14.4 [Range 11.3–17.8] mins	**Intra-observer: SPARS +**
	Ratingen					0.98 (0.85–0.99)	0.819 (0.802–0.838)				10.1 [Range 8.6–12.4] mins	**Inter-observer: SPARS +, mSvdHs +, Ratingen +**
						NR	0.869 (0.842–0.889)					**Measurement error: SPARS ± (MICs not defined)**
						NR						

(ROI: Radiographic Outcome Instrument; ICC: Intra-class correlation coeffiecient; OS: Original Steinbrocker; MS: Modified Steinbrocker; ML: Modified Larsen; NR: Not reported; MDC: Minimal Detectable Change; mTSS-A: modified Total Sharp Score-A; mTSS-B: modified Total

(B) Axial Radiographic Instruments

Study	ROI	No. of radiographs	Interval for Reliability	Observers stable? (n)	Test conditions stable?	Intra-Observer Reliability ICC (95% CI)	Inter-Observer Reliability ICC (95% CI)	Feasibility/ Mean time	Notes:
Lubrano 2009 PASRI	BASRI-T	73	2 weeks	Yes (3)	Yes			Subjective assessment of feasibility: BASRI quickest, PASRI = mSASSS	"Test-Retest Reliability" reported in paper, however radiographs were evaluated simultaneously by 3 observers to generate a consensus score BASRI-T: 0.97 (0.907–0.963) **Upper limit of 95% CI lower than ICC* mSASSS: 0.98 (0.966–0.987) No assessment of measurement properties undertaken as study rated Red from a methodological perspective due to the consensus interpretation of imaging.
	mSASSS								
Lubrano 2009	BASRI-T	77	10 months	Yes (3)	Yes			2 mins 5 mins	"Test-Retest Reliability" reported in paper, however radiographs were evaluated simultaneously by 3 observers to generate a consensus score BASRI-T: 0.97 (0.95–0.98) mSASSS: 0.98 (0.98–0.99) No assessment of measurement properties undertaken as study rated Red from a methodological perspective due to the consensus interpretation of imaging.
	mSASSS								
Biagioni 2014	mNYC	40 with PsA	2 weeks	Yes (4)	Yes	ICC: Two-way, mixed effects model, agreement,	0.67 (0.54–0.79)	1 hour training and e-handbook	**Intra-observer for PsA mNYC +**

(*continued on next page*)

Tables 2 (*Continued*)

					(B) Axial Radiographic Instruments				
Study	ROI	No. of radiographs	Interval for Reliability	Observers stable? (n)	Test conditions stable?	Intra-Observer Reliability ICC (95% CI)	Inter-Observer Reliability ICC (95% CI)	Feasibility/ Mean time	Notes:
	BASRI-S					single raters	0.52 (0.38–0.67)		**BASRI-S +**
	mSASSS					0.81 (0.73–0.88)	0.65 (0.49–0.80)		**mSASSS +**
	RASSS					0.77 (0.69–0.85)	0.68 (0.51–0.82)		**RASSS +**
	PASRI					0.91 (0.86–0.95)	0.88 (0.82–0.93)		**PASRI +**
						0.90 (0.85–0.95)			**Inter-observer for PsA**
						0.92 (0.87–0.95)			**mNYC**
	mNYC	18 with AS				0.91 (0.84–0.96)	0.80 (0.65–0.91)		**-BASRI-S**
	BASRI-S					0.96 (0.92–0.98)	0.86 (0.75–0.94)		**-mSASSS**
	mSASSS					0.98 (0.97–0.99)	0.86 (0.74–0.94)		**-RASSS**
	RASSS					0.96 (0.93–0.98)	0.75 (0.56–0.89)		**-PASRI +**
	PASRI					0.99 (0.97–0.99)	0.93 (0.86–0.97)		

Sharp Score-B; mSvdHs: modified Sharp van der Heidje score; ERO: Erosion; JSN: Joint Space Narrowing; SDC: Smallest Detectable Change; SDD: Smallest Detectable Difference; SEM: Standard Error of Mean; SPARS: Simplified Psoriatic Arthritis Radiographic Score; mSASSS: modified Stoke Ankylosing Spondylitis Spinal Score; BASRI-S: Bath Ankylosing Spondylitis Radiology Index – Spine; BASRI-T: Bath Ankylosing Spondylitis Radiology Index – Total; PASRI: Psoriatic Arthritis Spondylitis Radiology Index; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; mNYC: modified New York Criteria; PsA: Psoriatic Arthritis; AS: Ankylosing Spondylitis).

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 varies according to the study population and has not been defined for any instrument in these studies, 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 mSvdH score was conferred a GREEN rating for this measurement property. (Tables 3 and 4). [10]

Table 3
Construct validity and longitudinal construct validity.

Peripheral Radiographic Studies									
Study	Patients (n)	ROI	Domain of ROI	Radiographic Duration	Outcome Measured	Construct Validity againstRadiographic Comparator Instrument	Construct Validity againstNon-Radiographic Comparator Instrument	longitudinal construct validity	judgment/notes
Rahman 1998	68	ML Score vs. OS Score ML Score vs. MS Score	Total	2 years	Responsiveness analysis (regression slope) r^2 [95% CI] for Reader 1 and Reader 2			0.15 [0.06–0.24] and 0.09 [0.001–0.19] 1.1 [1.0–1.1] and 0.93 [0.8–1.0]	**Longitudinal construct validity:** **OS -** **MS and ML +** "Responsiveness was measured by plotting the change scores between baseline and 2 years for both methods, along with regression analysis. . .A slope near one would indicate that two methods being compared are reacting to changes to approximately the same degree" Radiographs selected to represent a sectrum of damage
Ravindran 2010	74	mTSS-A	Total	5.75 years (median)	Correlation at Baseline and Follow-up with: HAQ Clinical Joint Scores		r=0.29 and r=0.48 r=0.72 and r=0.81		**Construct validity: mTSS-A +** Strong association between mTSS-A and periarticular osteopaenia, bony proliferation, periostitis and bony ankyloses at baseline and follow-up ($p<0.001$). Soft-tissue swelling associated with increased rate of radiographic progression ($p<0.001$). Strong correlations between erosion and joint space narrowing at baseline and follow-up (r=0.83 and r=0.86 respectively)
Tillett 2014	50	MS Score mTSS-B mSvdHS Ratingen	 Total/ERO/JSN Total/ERO/JSN Total/Destruction/Proliferation	25 months (mean)	Standardised Response Mean			0.46 0.77/0.57/0.64 0.79/0.52/0.68 0.44/0.45/0.43	**Longitudinal construct validity:** **MS -** **mTSS-B +** **mSvdHs +** **Ratingen -** SDC was greater than the mean change over 2 years in all techniques. SRM of >0.8 suggests "high potential of detecting change" Known chronology.
Tillett 2016	50	ReXPsA	Total ERO JSN Proliferation	2.1 years	**Correlation (r) with:** mSvdHS/mTSS-B/Ratingen Score **Sensitivity compared to:** mSvdHS/mTSS-B/Ratingen Score **Correlation (r) with:** mSvdHS ERO/mTSS-B ERO/Ratingen Destruction **Sensitivity compared to:** mSvdHS ERO/mTSS-B ERO/Ratingen Destruction **Correlation (r) with:** mSvdHS JSN/mTSS-B JSN **Sensitivity compared to:** mSvdHS JSN/mTSS-B JSN **Correlation (r) with:** Ratingen proliferation **Sensitivity compared to:** Ratingen proliferation			 0.88/0.84/0.67 0.80/0.82/0.86 0.75/0.73/0.62 0.96/0.96/0.95 0.69/0.69 0.79/0.84 0.54 0.86	**Longitudinal construct validity:** **ReXPsA +** Analysis excluded small joints of the wrist Overall sensitivity to detect any change was 0.77. Correlations demonstrated with other radiographic measures should be interpreted with caution given the lack of validation outside of the cohort from which the tool was derived. Red on good methods checklist.
Kerschbaumer 2017 GO-REVEAL RCT All patients	363	mSvdHS	Total ERO JSN	104 weeks	Impact of each unit of increase of the radiographic score on HAQ-DI (Beta)			0.002 (0.001–0.003; p=0.005) 0.003 (0–0.005; p=0.019) 0.005 (0.002–0.007; p=0.001)	**Construct validity: mSvdHs +Longitudinal construct validity: mSvdHs ±**GEE Longitudinal Analysis – NOT predictive regression model. HAQ-DI at each visit was the dependent variable and radiographic damage were independent

(*continued on next page*)

Table 3 (*Continued*)

Peripheral Radiographic Studies									
Study	Patients (n)	ROI	Domain of ROI	Radiographic Duration	Outcome Measured	Construct Validity againstRadiographic Comparator Instrument	Construct Validity againstNon-Radiographic Comparator Instrument	longitudinal construct validity	judgment/notes
Kerschbaumer 2017 GO-REVEAL RCT Patients in cDAPSA LDA with abnormal HAQ >/=1 at baseline			Total		Ability of patients in the 3rr[d] Quartile of mSvdHS to achieve normal HAQ vs. patients in the 1st Quartile Absolute risk reduction in 1st Quartile of mSvdHS		Relative Risk 0.58 (0.35−0.96; p=0.029) ARR = 0.196		variables in separate models. Normal distribution with the identity link function was chosen as well as an autoregressive correlation matrix to account for within-subject correlations over time. The model was adjusted for DAPSA scores.Other GEE models reported in paper: • Using only the subgroup of patients with at least one visit in DAPSA remission. • Assessing impact of radiographic damage on the **absolute change** in HAQ-DI or **relative change** in HAQ-DI in Major Responders • Using the physical component of the SF-36 to assess the impact of radiographic damage on the SF36-PCS in patients in Remission • Using the physical component of the SF-36 to assess the impact on the responsiveness of SF36-PCS in Major responders. Conclusions: • Structural damage appeared to lead to functional disability, independent of disease activity. The effects appeared to be related more closely with JSN than ERO. • "Putting the estimate (Beta=0.002) of the remission model into clinical context, a patient in DAPSA remission with a mSvdHS of 10, 50, 100 or 150 would have a predicted 'residual' mean HAQ of 0.02, 0.1, 0.2 and 0.3 respectively"
Kerschbaumer 2017 GO-REVEAL RCT Remission patients			Total		Impact of each unit of increase of the radiographic score on other disease activity end-points: Patient Global Physician Global Patient global Pain TJC 68 SJC 66			0.02 (−0.0008−0.04; p=0.059) −0.007 (−0.026−0.012; p=0.484) 0.012 (−0.008−0.032; p=0.231) 0 (−0.002 − 0.002; p=0.902) 0 (−0.002−0.001; p=0.898)	
Kerschbaumer 2017 Validation cohort	160		Total ERO JSN	NR	Impact of each unit of increase of the radiographic score on HAQ-DI (Beta)			0.002 (0.003−0; p=0.021) vs. 0.003 (0.001−0.005; p<0.001) 0.002 (0.005- −0.001; p 0.172) vs. 0.007 (0.003−0.01; p<0.001) 0.003 (0.005−0.001; p=0.004) vs. 0.005 (0.003−0.007; p<0.001	External validation of data was conducted using a convenience sample of PsA patients from one centre's X-Ray database (n=206) with complete cDAPSA, HAQ and a corresponding radiographic assessment within 6 months of the clinical remission visit. 34.4% of the patients with radiographs achieved cDAPSA remission in the course of their disease.
Salaffi 2019	105	SPARS	Total	N/A					Construct Validity: SPARS +

(*continued on next page*)

Table 3 (*Continued*)

Peripheral Radiographic Studies

Study	Patients (n)	ROI	Domain of ROI	Radiographic Duration	Outcome Measured	Construct Validity againstRadiographic Comparator Instrument	Construct Validity againstNon-Radiographic Comparator Instrument	longitudinal construct validity	judgment/notes
					Correlation with mSvdHS Ratingen Score	r=0.926 ($p<0.0001$) r=0.904 ($p<0.0001$)			

Axial Radiographic Studies

Study	Patients (n)	ROI	Domain of ROI	Radiographic Duration	Outcome Measured	Construct Validity againstRadiographic Comparator Instrument	Construct Validity againstNon-Radiographic Comparator Instrument	Longitudinal Construct Validity	Comments
Chandran 2007	10 with AxPsA	mSASSS	Total	N/A	Correlations with spinal mobility measures in **AxPsA** vs. AS Occiput-to-wall Tragus-to-wall Cervical rotation Chest expansion Modified Schober Domjan Lateral Spinal Flexion INSPIRE Lateral Spinal Flexion		**0.86 [0.50, 0.97] $p<0.01$** vs. 0.84 [0.40, 0.97] $p<0.01$ **0.83 [0.41, 0.96] $p<0.01$** vs. 0.75 [0.17, 0.94] $p<0.05$ **−0.83 (−0.96, −0.41] $p<0.01$** vs. −0.11 [−0.72, 0.60] **−0.61 [−0.90, 0.03] NS** vs. −0.46 [−0.86, −0.30] NS **−0.59 [−0.89, 0.06] NS** vs. −0.82 [−0.96, −0.35] $p<0.05$ **−0.66 [−0.91, −0.05] $p<0.01$** vs. −0.77 [−0.95, −0.22] $p<0.01$ **−0.60 [−0.89, 0.05] NS** vs. −0.87 [−0.97, −0.47] $p<0.001$		**Construct Validity: mSASSS ±** BASRI-S measured but correlations not reported Spinal mobility measurements used were a median of scores from 10 assessors.
Lubrano 2009 PASRI	73	BASRI-T mSASSS \PASRI	Total	N/A	Correlation (Spearman rho) with: PASRI BASMI/BASFI/RLQD Cervical rotation Tragus to wall Occiput to wall Chest expansion Chest expansion (nipple) Modified Schober's Finger to floor Intermalleolar distance PASRI BASMI/BASFI/RLQD Cervical rotation Tragus to wall Occiput to wall Chest expansion Chest expansion (nipple) Modified Schober's Finger to floor Intermalleolar distance BASRI-T mSASSS BASMI/BASFI/RLQD Cervical rotation Tragus to wall Occiput to wall Chest expansion Chest expansion (nipple) Modified Schober's Finger to floor Intermalleolar distance	 0.80 (p = 0.0001) 0.94 (p = 0.0001) 0.80 (p = 0.0001) 0.94 (p = 0.0001)	 0.65 (p = 0.01)/0.51 (p = 0.01)/0.46 (p = 0.01) 0.41 (p=0.05) 0.46 (p=0.05) 0.38 (p=0.05) −0.34 (p=0.01) −0.61 (p=0.01) −0.54 (p=0.01) 0.44 (p=0.01) −0.69 (p=0.01) 0.46 NS/0.36 NS/0.22 NS −0.15 (NS) 0.36 (NS) 0.30 (NS) 0 (NS) −0.33 (NS) −0.49 (p=0.01) 0.42 (NS) −0.01 (NS) 0.60 (p = 0.01)/0.50 (p = 0.01)/0.47 (p = 0.01) −0.30 (NS) 0.40 (p=0.05) 0.41 (NS) −0.15 (NS) −0.28 (NS) −0.49 (p=0.01) 0.54 (p=0.01) −0.36 (p=0.01)		**Construct Validity: mSASSS - BASRI-T + PASRI +** Wanders' adjustment utilised for mSASSS: If >3 scoring sites were missing, the radiographs were excluded. If 3 or fewer sites were missing, the mean if the other scoring sites were used as a substitute for the missing sites (n=13
Lubrano 2009	77	BASRI-T/ mSASSS BASRI-T	Total	N/A	Correlations (rho) with: BASMI Cervical rotation Tragus to wall	 0.855, $p<0.001$	 0.47 ($p<0.01$) / 0.39 ($p<0.05$) −0.49 ($p<0.01$) / −0.41 ($p<0.01$) 0.34 ($p<0.01$) / 0.31 ($p<0.01$)		**Construct Validity: mSASSS ± BASRI-*T* ±** Wanders' adjustment

(*continued on next page*)

Table 3 (*Continued*)

Axial Radiographic Studies									
Study	Patients (n)	ROI	Domain of ROI	Radiographic Duration	Outcome Measured	Construct Validity againstRadiographic Comparator Instrument	Construct Validity againstNon-Radiographic Comparator Instrument	Longitudinal Construct Validity	Comments
					Occiput to wall		0.49 ($p<0.01$) / 0.42 ($p<0.01$)		utilised for mSASSS: If >3 scoring sites were missing, the radiographs were excluded. If 3 or fewer sites were missing, the mean if the other scoring sites were used as a substitute for the missing sites (n = 13
					Chest expansion		−0.27 ($p<0.05$) / −0.15 NS		
					Chest expansion (nipple)		−0.34 ($p<0.01$) / −0.26 ($p<0.05$)		
					Modified Schober		−0.24 ($p<0.05$) / −0.34 ($p<0.01$)		
					Schober		−0.22 NS / −0.32 ($p<0.01$)		
					Finger to floor		0.22 NS / 0.37 ($p<0.01$)		
					Intermallolar distance		−0.20 NS / −0.09 NS		
					BASFI		0.14 NS / 0.010 NS		
					RLDQ		0.24 ($p<0.001$) / 0.12 NS		
					HAQ		0.00 NS / 0.05 NS		
					mSASSS				
Ibrahim 2017	105	BASRI-S	Total	≥2 years	% of Patients who progressed			29%	**Longitudinal Construct Validity:**
		mSASSS						25%	**BASRI-*S* ±**
		RASSS						23%	**mSASSS ±**
		PASRI						32%	**RASSS ±**
		Independent assessor						24%	**PASRI ±**
		BASRI-S			Sensitivity in detecting 'true' progression with a score increase of >/1 vs. ≥2			≥1 0.48 (0.28−0.68) ≥2 0.12 (0.00−0.25)	All instruments were comparable to the binary outcome (PASRI performed the best) in terms of detecting progression, but the sensitivity of a slight increase in score was poor.
		mSASSS						≥1 0.52 (0.32−0.72) ≥2 0.32 (0.14−0.50)	
		RASSS						≥1 0.44 (0.25−0.65) ≥2 0.32 (0.14−0.50)	
		PASRI						≥1 0.52 (0.32−0.72) ≥2 0.48 (0.28−0.68)	
		BASRI-S			Specificity in detecting 'true' progression with a score of >/=1 vs. ≥2			≥1 0.78 (0.67−0.86) ≥2 0.96 (0.92−1.00)	
		mSASSS						≥1 0.84 (0.73−0.91) ≥2 0.88 (0.82−0.96)	
		RASSS						≥1 0.84 (0.73−0.91) ≥2 0.88 (0.80−0.95)	
		PASRI						≥1 0.74 (0.63−0.83) ≥2 0.84 (0.76−0.92)	
		BASRI-S			Logistic regression analyses (age, OA, DISH) to determine OR (95% CI) for identifying 'true progression' as determined by independent assessor per unit increase in instrument score			3.00 (1.15−7.82), p = 0.024	
		mSASSS						5.27 (1.92−14.48), p = 0.001	
		RASSS						3.70 (1.32−10.36) p = 0.013	
		PASRI						3.06 (1.18−7.95) p = 0.022	

(MS: Modified Steinbrocker; OS: Original Steinbrocker; ML: Modified Larsen; mTSS-A: modified Total Sharp Score-A; mTSS-B: modified Total Sharp Score-B; mSvdHs: modified Sharp van der Heidje score; ReXPSA: Reductive X-Ray Score for Psoriatic Arthritis; SPARS: Simplified Psoriatic Arthritis Radiographic Score; ERO: Erosion; JSN: Joint Space Narrowing; HAQ-DI: Health Assessment Questionnaire-Disability Index; SF-36: Short Form 36 Health Survey; SF-36: PCS: Short Form 36 Health Survey: Physical Component Score; c-DAPSA: Clinical Disease Activity Index for Psoriatic Arthritis; DAPSA: Disease Activity Index for Psoriatic Arthritis; LDA: Low Disease Activity; ARR: Absolute Risk Reduction; RCT: Randomised Controlled Trial; TJC: Tender Joint Count; SJC: Swollen Joint Count; PsA: Psoriatic Arthritis; mSASSS: modified Stoke Ankylosing Spondylitis Spinal Score; BASRI-S: Bath Ankylosing Spondylitis Radiology Index − Spine; BASRI-T: Bath Ankylosing Spondylitis Radiology Index − Total; PASRI: Psoriatic Arthritis Spondylitis Radiology Index; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; mNYC: modified New York Criteria; OA: Osteoarthritis; DISH: Diffuse Idiopathic Skeletal Hyperostosis; AS: Ankylosing Spondylitis; AxPsA: Axial Psoriatic Arthritis; BASMI: Bath Ankylosing Spondylitis Metrology Index; BASFI: Bath Ankylosing Spondylitis Functional Index; RLQD: Revised Leeds Disability Questionnaire.

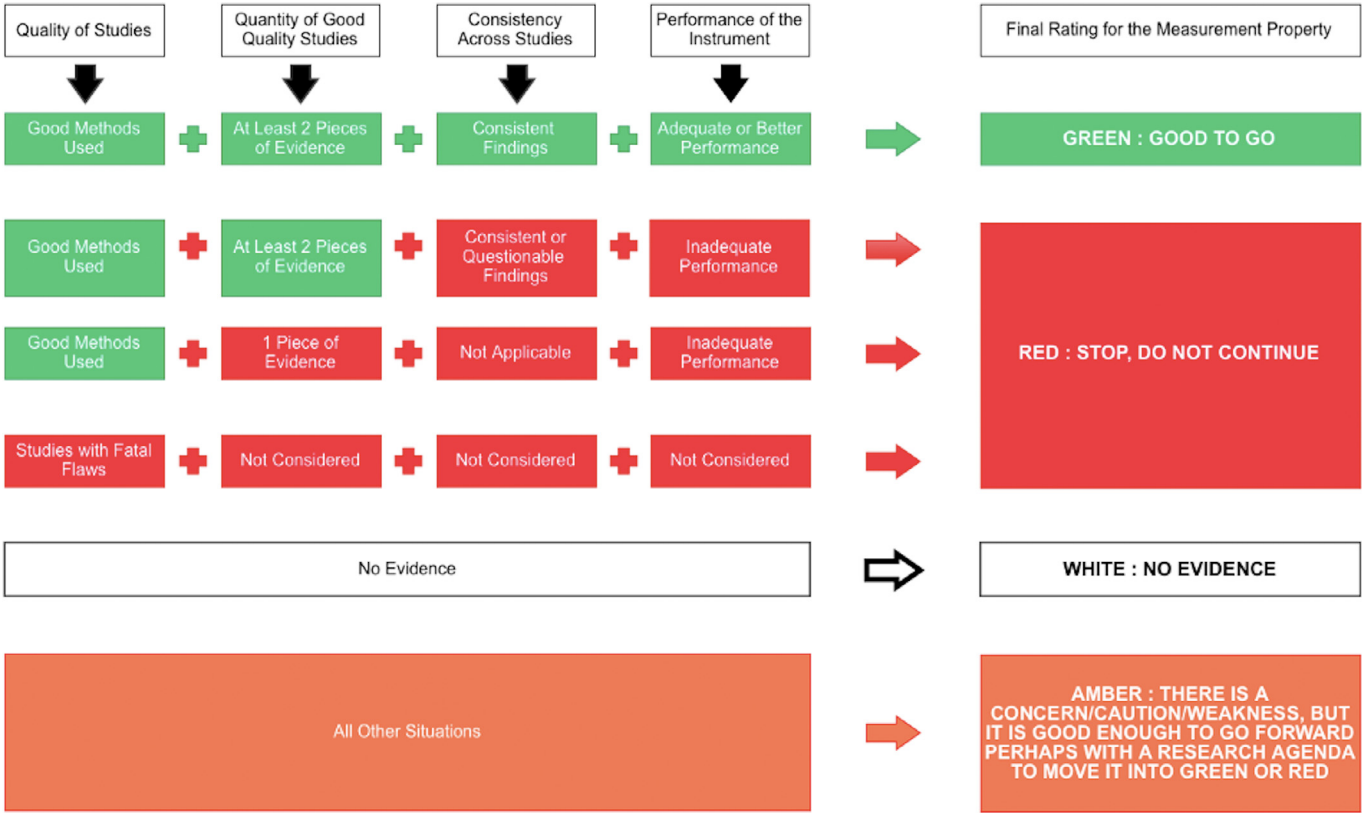

Fig. 1. Synthesis of evidence to generate a final RED/AMBER/WHITE/GREEN (RAWG) Rating.

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) fulfilled CASPAR, 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 min 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

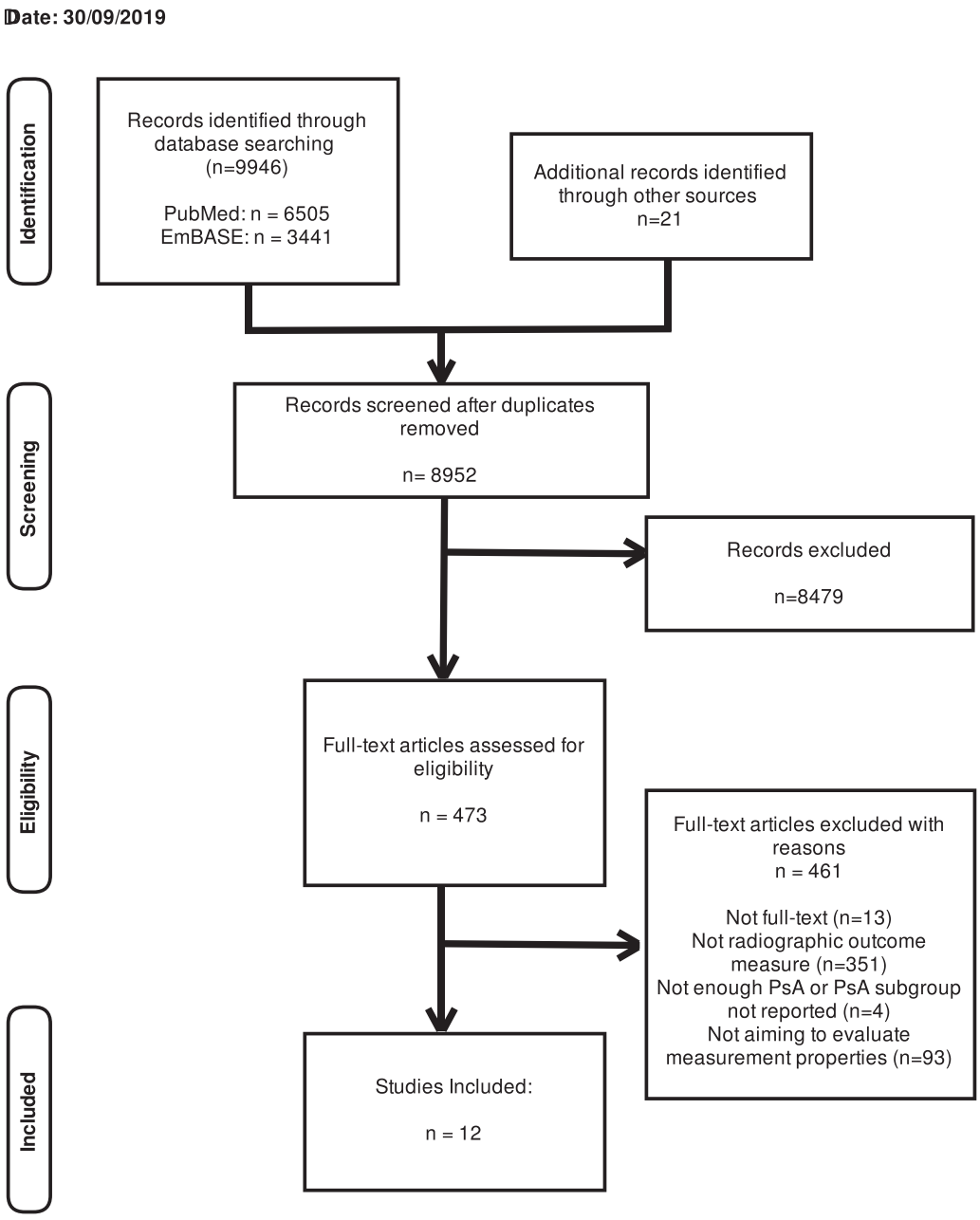

Fig. 2. Prisma flow diagram.

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 $\geq$1 with each instrument was comparable, sensitivity was highest with the mSASSS and PASRI, and the specificity was

Table 4
Summary of Measurement Properties – Peripheral ROIs.

Author/Year	Domain match	Feasibility	Truth construct validity	Discrimination						
				Reliability				Responsiveness		
				Inter-rater	Intra-rater	Test-retest	Measurement error	Longitudinal construct validity	Clinical trial discrimination	Thresholds of meaning
	Original Steinbrocker Score									
Rahman 1998				Amber +	Amber +			Amber -		
Total Available Studies				1	1			1		
Studies for Synthesis				1	1			1		
Rating (RAGW)	WHITE	WHITE	WHITE	AMBER	AMBER	WHITE	WHITE	AMBER	WHITE	WHITE
Modified Steinbrocker Score										
Rahman 1998				Amber +	Amber +			Amber ±		
Tillett 2014				Amber -	Amber +		Green ±	Amber -		
Total Available Studies				2	2		1	2		
Studies for Synthesis				2	2		1	2		
Rating (RAGW)	WHITE	WHITE	WHITE	AMBER	AMBER	WHITE	AMBER	AMBER	WHITE	WHITE
Modified Larsen Score										
Rahman 1998				Amber +	Amber +			Amber +		
Total Available Studies				1	1			1		
Studies for Synthesis				1	1			1		
Rating (RAGW)	WHITE	WHITE	WHITE	AMBER	AMBER	WHITE	WHITE	AMBER	WHITE	WHITE
Ratingen Score										
Wassenberg 2001				Amber ±	Amber ±		Green ±			
Tillett 2014				Amber +	Amber +		Green ±	Amber -		
Salaffi 2019			Green +	Amber +						
Total Available Studies			1	3	2		2	1		
Studies for Synthesis			1	3	2		2	1		
Rating (RAGW)	WHITE	WHITE	AMBER	AMBER	AMBER	WHITE	AMBER	AMBER	WHITE	WHITE
Modified Total Sharp Score - A										
Ravindran 2010			Amber ±	Amber +	Amber +					
Total Available Studies			1	1	1					
Studies for Synthesis			1	1	1					
Rating (RAGW)	WHITE	WHITE	AMBER	AMBER	AMBER	WHITE	WHITE	WHITE	WHITE	WHITE
Modified Total Sharp Score - B										
Tillett 2014				Amber +	Amber +		Green ±	Amber +		
Total Available Studies				1	1		1	1		
Studies for Synthesis				1	1		1	1		
Rating (RAGW)	WHITE	WHITE	WHITE	AMBER	AMBER	WHITE	AMBER	AMBER	WHITE	WHITE
Modified Sharp van der Heidje Score										
Tillett 2014				Amber +	Amber +		Green ±	Amber +		
Kerschbaumer 2017			Green +					Green ±		
Salaffi 2019			Green +	Amber +						
Total Available Studies			2	2	1		1	2		
Studies for Synthesis			2	2	1		1	2		
Rating (RAGW)	WHITE	WHITE	GREEN	AMBER	AMBER	WHITE	AMBER	AMBER	WHITE	WHITE
Reductive X-Ray Score for Psoriatic Arthritis										
Tillett 2016								Red +		
Total Available Studies								1		
Studies for Synthesis								0		
Rating (RAGW)	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE
Simplified Psoriatic Arthritis Radiographic Score										
Salaffi 2019			Green +	Amber +	Amber +					
Total Available Studies			1	1	1					
Studies for Synthesis			1	1	1					
Rating (RAGW)	WHITE	WHITE	AMBER	AMBER	AMBER	WHITE	WHITE	WHITE	WHITE	WHITE

Table 5
Summary of measurement properties for axial ROIs.

Axial Psoriatic Arthritis Definition 1 (CASPAR + Spinal inflammatory pain and/or radiologic axial involvement										
Author/Year	Domain Match	Feasibility	Truth Construct Validity	Discrimination						
				Reliability				Responsiveness		
				Inter-rater	Intra-rater	Test-Retest	Measurement Error	Longitudinal Construct Validity	Clinical Trial Discrimination	Thresholds of Meaning
Modified Stoke Ankylosing Spondylitis Score										
Lubrano 2009			Amber ±			Red				
Lubrano PASRI 2009			Amber ±			Red				
Total Available Studies			2			2				
Studies for Synthesis			2			0				
Rating (RAGW)	WHITE	WHITE	AMBER	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE
Bath Ankylosing Spondylitis Radiology Index - Total										
Lubrano 2009			Amber ±			Red				
Lubrano PASRI 2009			Amber ±			Red				
Total Available Studies			2			2				
Studies for Synthesis			2			0				
Rating (RAGW)	WHITE	WHITE	AMBER	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE
Psoriatic Arthritis Spondylitis Radiology Index										
Lubrano PASRI 2009			Amber ±							
Total Available Studies			1							
Studies for Synthesis			1							
Rating (RAGW)	WHITE	WHITE	AMBER	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE	WHITE

Axial Psoriatic Arthritis Definition 2 (CASPAR + At least unilateral grade II sacroiliitis AND IBP or Restricted Spinal Mobility)

Author/Year	Domain Match	Feasibility	Truth Construct Validity	Discrimination						
				Reliability				Responsiveness		
				Inter-rater	Intra-rater	Test-Retest	Measurement Error	Longitudinal Construct Validity	Clinical Trial Discrimination	Thresholds of Meaning
Modified Stoke Ankylosing Spondylitis Score										
Chandran 2007			Amber +							
Biagioni 2014				Green -	Green +					
Ibrahim 2017								Amber ±		
Total Available Studies			1	1	1			1		
Studies for Synthesis			1	1	1			1		
Rating (RAGW)	WHITE	WHITE	AMBER	RED	AMBER	WHITE	WHITE	AMBER	WHITE	WHITE
Bath Ankylosing Spondylitis Radiology Index - Spine										
Biagioni 2014				Green -	Green +					
Ibrahim 2017								Amber ±		
Total Available Studies				1	1			1		
Studies for Synthesis				1	1			1		
Rating (RAGW)	WHITE	WHITE	WHITE	RED	AMBER	WHITE	WHITE	AMBER	WHITE	WHITE
Psoriatic Arthritis Spondylitis Radiology Index										
Biagioni 2014				Green +	Green +					
Ibrahim 2017								Amber ±		
Total Available Studies				1	1			1		
Studies for Synthesis				1	1			1		
Rating (RAGW)	WHITE	WHITE	WHITE	AMBER	AMBER	WHITE	WHITE	AMBER	WHITE	WHITE
Modified New York Criteria										
Biagioni 2014				Green -	Green +					
Total Available Studies				1	1					

(*continued on next page*)

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 was 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 oSteinbrocker 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 oSteinbrocker 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. The mSteinbrocker however, has demonstrated comparable change over time when compared to the Ratingen and mSvdH score in an observational study of a single patient with a baseline mSvdH score of 59 (Range 0–528). [36]

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. [6,37] 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. [38] A number of RCTs have assessed the yield of assessing for proliferative features

Table 5 (*Continued*)

Author/Year	Domain Match	Feasibility	Truth Construct Validity	Discrimination							
				Reliability				Responsiveness			
				Inter-rater	Intra-rater	Test-Retest	Measurement Error	Longitudinal Construct Validity	Clinical Trial Discrimination	Thresholds of Meaning
Studies for Synthesis Rating (RAGW)	WHITE	WHITE	WHITE	1 RED	1 AMBER	WHITE	WHITE	WHITE	WHITE	WHITE
Radiographic Ankylosing Spondylitis Spinal Score										
Biagioni 2014				Green -	Green +			Amber ±		
Ibrahim 2017								1		
Total Available Studies				1	1			1		
Studies for Synthesis Rating (RAGW)	WHITE	WHITE	WHITE	RED	1 AMBER	WHITE	WHITE	1 AMBER	WHITE	WHITE

such as osteitis and ankyloses, and have not noted a significant progression in these features over time nor a significant difference between treatment arms. [14,15,21,22,27,39–42] 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 drug efficacy. [14–29] Furthermore, it is important to note that the mSvdH score 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. OMERACT defines thresholds of meaning as "the degree to which one can assign an easily understood meaning to the scores from an instrument", which may include a patient acceptable symptom state or a minimum important improvement [4]. The minimum important improvement or change of an outcome instrument can be assessed in a number of ways, including the minimal clinically important difference (MCID) or the minimally important difference (MID). The MCID is typically utilised anchor-based methods and is employed to assess thresholds of meaning for patient-reported outcomes [43]. Determining the degree of radiographic change that is likely to be perceived as 'clinically significant' may not be feasible given the non-linear relationship between radiographic damage and function. This is likely to vary significantly from patient to patient depending on their baseline damage, co-existing disease activity, and the joints affected. Furthermore, it may not be possible for patients to indicate a difference that is important to them other than "progression" or "no progression". The MID is a more appropriate measure for the assessment of thresholds of meaning in imaging instruments. The MID relies on distribution-based methods to assess the measurement error within particular population, and will therefore vary between populations. In RCTs, this is typically assessed as 'any progression' or 'any progression above the smallest detectable difference within the study'. Patients rank prevention of damage as a highly important outcome and there is an argument that any progression or damage accumulation may be important even if there is no discernible impact on function. [44]

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, and determining clinical trial discrimination and thresholds of meaning. [42,45] The OMERACT working group will proceed to ascertaining consensus regarding domain match and feasibility prior to 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'. [46–50] 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'. [48,51]

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. [38,49,52,53] 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 o 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, what constitutes an acceptable interobserver reliability and whether a reliability exercise should to be undertaken prior to formal scoring, 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 or an increase above the measurement error) and the imputation of missing data. The 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 clinical trial discrimination and thresholds of meaning. [28,54] 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 possible that this will progressively shift to earlier disease in future trials. [14,15,21,22,27,39–42] 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, 55]

A further knowledge gap is the discriminative capacity of radiographic instruments to differentiate changes related to PsA from those related to 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. [56–58] 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 in our analysis, 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. This systematic review provides a substrate on which future recommendations can be made.

Declaration of Competing Interest

Antony A, Holland R, D'Agostino MA, Maksymomwych W.P., Bertheussen H, Schick L, Goel N, Orbai AM, Højgaard P, Coates L, Strand V, Christensen R, Leung YY and Tillett W have no conflicts to declare.

Dr. Ogdie reports personal fees from Abbvie, grants and personal fees from Amgen, personal fees from BMS, personal fees from Celgene, personal fees from Corrona, personal fees from Janssen, personal fees from Lilly, grants, personal fees and other from Novartis, grants and personal fees from Pfizer, outside the submitted work; Husband receives royalties from Novartis.

Dr. Gladman reports grants and personal fees from Abbvie, grants and personal fees from Amgen, personal fees from BMS, grants and personal fees from Celgene, grants and personal fees from Eli Lilly, personal fees from Gilead, personal fees from Galapagos, grants and personal fees from Janssen, grants and personal fees from Novartis, grants and personal fees from Pfizer, grants and personal fees from UCB, outside the submitted work;

Dr. Mease reports grants and personal fees from AbbVie, grants and personal fees from Amgen, grants and personal fees from Bristol Myers Squibb, personal fees from Boehringer Ingelheim, grants and personal fees from Celgene, personal fees from Galapagos, personal fees from Genentech, personal fees from Gilead, personal fees from GlaxoSmithKline, grants and personal fees from Janssen, grants and personal fees from Lilly, grants and personal fees from Novartis, grants and personal fees from Pfizer, grants and personal fees from Sun, grants and personal fees from UCB, outside the submitted work.

Dr. Leung reports personal fees from AbbVie, Novartis, Elli Lilly and Janssen outside if the submitted work.

Dr Tillett reports grants from Abbvie, Celgene, Janssen, Lilly, and personal fees from Abbvie, Amgen, Celgene, Lilly, Janssen, MSD, Novartis, Pfizer and UCB, outside the submitted work.

All authors are members of OMERACT, an organization that develops and validates outcome measures in rheumatology randomized

controlled trials and longitudinal observational studies and receives arms-length funding from 36 sponsors.

Professor R. Christensen acknowledges that the Parker Institute, Bispebjerg and Frederiksberg Hospital is supported by a core grant from the Oak Foundation (OCAY-18-774-OFIL). None of the funding sources had any influence on the study design, on data collection, data synthesis, data interpretation, writing the report, or the decision to submit the manuscript for publication.

Acknowledgment

The authors thank Lara Maxwell and Dorcas Beaton from the OMERACT Technical Advisory Group for their assistance in reviewing the data extraction and synthesis.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.semarthrit.2021.01.008.

References

1. Orbai AM, 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.
2. Beaton DE, et al. Instrument selection using the OMERACT Filter 2.1: the OMERACT methodology. J Rheumatol 2019;46(8):1028–35.
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):654–65.
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):1033–41.
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):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 2014;66(2):311–7.
8. Rahman P, et al. Radiological assessment in psoriatic arthritis. Br J Rheumatol 1998;37(7):760–5.
9. Wassenberg S, et al. A method to score radiographic change in psoriatic arthritis. Z Rheumatol 2001;60(3):156–66.
10. Ravindran J, et al. A modified Sharp score demonstrates disease progression in established psoriatic arthritis. Arthritis Care Res 2010;62(1):86–91.
11. Mokkink LB, D AM. Protocol for performing a systematic review on imaging techniques. 2017.
12. Prinsen CAC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27(5):1147–57.
13. Kerschbaumer A, et al. The effects of structural damage on functional disability in psoriatic arthritis. Ann Rheum Dis 2017;76(12):2038–45.
14. Mease PJ, et al. Etanercept treatment of psoriatic arthritis: safety, efficacy, and effect on disease progression. Arthritis Rheum 2004;50(7):2264–72.
15. Mease PJ, 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):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):890–7.
17. Mease PJ, et al. Secukinumab inhibition of Interleukin-17A in patients with Psoriatic Arthritis. N Engl J Med 2015;373(14):1329–39.
18. Mease PJ, 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):1112–24.
19. Fraser AD, 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):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):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):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):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):2151–61.
24. Mease PJ, 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):1550–8.
25. Mease PJ, 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):79–87.
26. Mease P, et al. Tofacitinib or Adalimumab versus Placebo for Psoriatic Arthritis. N Engl J Med 2017;377(16):1537–50.
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):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):233–7.
29. Coates LC, 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):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):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):1006–11.
32. Biagioni BJ, et al. Reliability of radiographic scoring methods in axial psoriatic arthritis. Arthritis Care Res 2014;66(9):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):2463–5.
34. Ibrahim A, et al. Sensitivity and specificity of radiographic scoring instruments for detecting change in Axial Psoriatic Arthritis. Arthritis Care Res 2017;69(11):1700–5.
35. Siannis F, et al. Clinical and radiological damage in psoriatic arthritis. Ann Rheum Dis 2006;65(4):478–81.
36. Eder L, Chandran V, Gladman DD. Repair of radiographic joint damage following treatment with etanercept in psoriatic arthritis is demonstrable by 3 radiographic methods. J Rheumatol 2011;38(6):1066–70.
37. 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).
38. 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 2019;58(2):269–73.
39. Mease PJ, et al. Continued inhibition of radiographic progression in patients with psoriatic arthritis following 2 years of treatment with etanercept. J Rheumatol 2006;33(4):712–21.
40. 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):2254–9.
41. Antoni CE, 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):869–76.
42. 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):595–602.
43. Strand V, 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.
44. Dures E, et al. Important treatment outcomes for patients with Psoriatic Arthritis: a multisite qualitative study. Patient 2017;10(4):455–62.
45. 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):1689–94.
46. Yap KS, 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):1573–7.
47. Fernandez-Sueiro JL, et al. Validity of the bath ankylosing spondylitis disease activity index for the evaluation of disease activity in axial psoriatic arthritis. Arthritis Care Res 2010;62(1):78–85.
48. Jadon DR, et al. Axial Disease in Psoriatic Arthritis study: defining the clinical and radiographic phenotype of psoriatic spondyloarthritis. Ann Rheum Dis 2017;76(4):701–7.
49. Queiro R, et al. Clinically asymptomatic axial disease in psoriatic spondyloarthropathy. A retrospective study. Clin Rheumatol 2002;21(1):10–3.
50. Feld J, et al. Axial disease in psoriatic arthritis and ankylosing spondylitis: a critical comparison. Nat Rev Rheumatol 2018;14(6):363–71.

51. Helliwell PS, Hickling P, Wright V. 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):135–40.
52. Chandran V, et al. Axial psoriatic arthritis: update on a longterm prospective study. J Rheumatol 2009;36(12):2744–50.
53. 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):1057–62.
54. 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):49–60.
55. Helliwell PS, et al. Comparing Psoriatic Arthritis low-field magnetic resonance imaging, ultrasound, and clinical outcomes: data from the TICOPA trial. J Rheumatol 2020;47(9):1338–43.
56. 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):2034–8.
57. 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):2473–9.
58. 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):1394–400.