Comparing the construct validity of measurement instruments for pain and stiffness in patients with axial spondyloartwhritis: cross-sectional analysis in the OASIS cohort

Capelusnik, D.; Nikiphorou, E.; Boonen, A.; Christensen, R.; Heijde, D. van der; Landewe, R.; ... ; Ramiro, S.

RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Comparing the construct validity of measurement instruments for pain and stiffness in patients with axial spondyloarthritis: cross-sectional analysis in the OASIS cohort

Dafne Capelusnik ,[1,2] Elena Nikiphorou ,[3,4] Annelies Boonen ,[1,5] Robin Christensen,[6,7] Désirée van der Heijde ,[8] Robert Landewé ,[9,10] Astrid van Tubergen ,[1,5] Sofia Ramiro [8,11]

To cite: Capelusnik D, Nikiphorou E, Boonen A, *et al*. Comparing the construct validity of measurement instruments for pain and stiffness in patients with axial spondyloarthritis: cross-sectional analysis in the OASIS cohort. *RMD Open* 2024;**10**:e004775. doi:10.1136/rmdopen-2024-004775

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/rmdopen-2024-004775).

Received 18 July 2024
Accepted 4 November 2024

For numbered affiliations see end of article.

Correspondence to
Dr Sofia Ramiro;
sofiaramiro@gmail.com

ABSTRACT

Objectives To compare the construct validity, including discrimination between known groups, of three pain and three morning stiffness (MS) measurement instruments.

Methods Patients with radiographic axial spondyloarthritis with 8-year data from the Outcome in Ankylosing Spondylitis International Study cohort were assessed cross-sectionally. Three instruments for pain and three for MS, all self-reported and scored 0–10, were compared. Construct validity was evaluated by testing (1) hypothesis of correlations' strength and (2) discrimination between known groups using standardised mean differences (SMD) across external constructs. Influence of contextual factors (CFs) on SMDs was investigated.

Results Of 85 patients, mean age was 54 (SD 11), mean symptom duration 31 (11) years, 71% males. All six instruments showed a good construct validity by fulfilling >75% of the hypotheses for the strength of correlation. Neck/back/hip pain (Bath Ankylosing Spondylitis Disease Activity Index-Question 2, BASDAI-Q2) and total back pain had higher SMDs compared with back pain at night across all between-group comparisons, with BASDAI-Q2 performing mostly slightly better (eg, SMD for external construct Axial Spondyloarthritis Disease Activity Score (ASDAS; ≥2.1 vs <2.1): 1.87 (BASDAI-Q2) vs 1.56 (total back pain) vs 1.07 (back pain at night)). MS-severity and severity/duration had higher SMDs across all external constructs (with MS-severity slightly better), while MS-duration performed worse (eg, SMD external construct ASDAS: 1.51 (MS-severity) and 1.39 (MS-severity/duration) vs 1.16 (MS-duration)). Influence of CFs on known group discrimination was limited.

Conclusions The recommended Assessment of SpondyloArthritis international Society Core Outcome Set (ASAS-COS) pain measurement instrument total back pain BASDAI-Q2 has the best known group discrimination. For MS, the ASAS-COS stiffness measure (MS-severity/duration) performs well although MS-severity even slightly better. Known group discrimination is overall stable across CFs.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ In the recently updated Assessment of SpondyloArthritis international Society Core Outcome Set for axial spondyloarthritis (axSpA), pain and stiffness remain mandatory domains to assess outcomes in clinical studies, with total back pain (pain in neck, back and hips) from the Bath Ankylosing Spondylitis Disease Activity Index-Question 2 (BASDAI-Q2), and the average between severity and duration of morning stiffness (MS) from BASDAI-Questions 5/6 (Q5/6) as the recommended instruments. This instrument selection was based on the assessment of psychometric properties, and among them, construct validity, which was assessed by the correlation of strength with external constructs. Nevertheless, this approach does not allow a comparison across different instruments.

WHAT THIS STUDY ADDS

⇒ Through the comparison of the discrimination between known groups, we demonstrated that total back pain (BASDAI-Q2) has better performance than back pain at night in radiographic axSpA; and that MS-severity (BASDAI-Q5) and MS-severity/duration (BASDAI-Q5/6) have better known group discrimination than MS-duration (BASDAI-Q6). Overall, the discriminatory capacity of pain and MS instruments is stable across contextual factors.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Total back pain from BASDAI-Q2 should be used for the back pain assessment. For MS, the severity or the average of severity/duration (but not duration itself) should be used.

INTRODUCTION

Pain and stiffness are two of the main symptoms of axial spondyloarthritis (axSpA). Also,

in the recently updated Assessment of SpondyloArthritis international Society Core Outcome Set (ASAS-COS) for axSpA, pain and stiffness remain mandatory domains to assess outcomes in clinical studies.[1] In the domain pain, two instruments were evaluated as candidates for the ASAS-COS: the numerical rating scale (NRS) for total back pain (pain in neck, back and hips), obtained from Question 2 from the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI-Q2), and the NRS for back pain at night in the past week.[2 3] The extensive literature review that informed the selection of the preferred instrument by ASAS showed that both instruments performed similarly for all psychometric properties but results for construct validity were inconsistent for total back pain (BASDAI-Q2) while good for back pain at night.[1] However, as a substantial number of persons with axSpA have no night pain (floor effect) and as total back pain was more frequently included in clinical studies, the latter was chosen to be included in the COS. In the domain morning stiffness, three instruments, all stemming from BASDAI,[2] were considered: morning stiffness-severity (BASDAI-Q5), morning stiffness-duration (BASDAI-Q6) and morning stiffness-severity/duration (BASDAI-Q5/6). Following the literature review and discussion, the morning stiffness-severity/duration was selected as the preferred outcome instrument as it fulfilled criteria for all psychometric properties, except for the absence of thresholds of meaning. It was, however, striking that only one of 12 studies reporting construct validity assessed the three instruments and only two external constructs were included, indicating limited evidence of a preferred instrument regarding construct.

The Outcome Measures in Rheumatology (OMERACT) filter for instrument selection is based on the three pillars of OMERACT: truth, discrimination and feasibility.[4] Truth is defined as the match of the instrument with the target domain and reflects whether the instrument has the right content for the experience of that domain in the intended target population (face and content validity). Truth also refers to the numbers/scores obtained from the instrument and whether they are in line with what would have been expected from our knowledge of the outcome domain intended to be measured (construct validity). The term 'discrimination' in the filter refers to the extent the instrument can discriminate between situations of interest, and includes reliability, responsiveness (longitudinal construct validity), discrimination in clinical trials and thresholds of meaning. Lastly, feasibility provides evidence to determine whether it is practical to use a given instrument.

Construct validity is assessed by the extent to which the scores on the instrument are consistent with hypotheses with regard to relationships with other instruments, or differences between relevant groups with known differences. Although the latter is mentioned in both the OMERACT filter and the Consensus-based Standards for the Selection of Health Measurement Instruments taxonomy,[5 6] no guidance is provided on how to address it.[7] In addition to the 'hypothesis testing based on correlations', 'known-group discrimination' provides a more comprehensive assessment on construct validity, especially when several measurement instruments on the same domain are tested, by allowing the comparison across them; for instance, how much an instrument of pain/morning stiffness discriminates or not between different states of a relevant domain like disease activity, fatigue or functional ability. This allows to assess each instrument property and to make a true comparison across the instruments.

More recently, the role of contextual factors in outcome measurement has received attention. Contextual factors are defined as variables that are not an outcome of the study but need to be recognised (and measured) to understand the study results. Three methodological types are distinguished: effect modifying, measurement affecting and outcome explaining. Although effect-modifying contextual factors are relevant to intervention studies, exploration of the role of contextual factors on the validity of the instrument itself or on the score of the outcome provides relevant insight when testing construct validity. In other words, it is essential to understand whether the 'construct' of the measurement instrument is similar across groups of persons that differ in context such as age, gender or educational level. For example, stiffness might be related to age (influence by age per se) and this may therefore affect measurement properties and/or interpretation of the scores.[8 9]

Acknowledging the above challenges, the objective of this project was to get further insight into construct validity, including known group discrimination, of three outcome measures used to assess pain and morning stiffness, and further validate the choices made in the axSpA COS. In addition, we aimed to evaluate the robustness of the selected instruments across contextual factors.

METHODS

Patient recruitment

Data from the Outcome in Ankylosing Spondylitis International Study (OASIS) cohort were used. OASIS is a prevalence cohort, started in 1996 and including 217 consecutive patients with radiographic axSpA (r-axSpA) from the Netherlands, Belgium and France.[10] Clinical and radiographic data were collected at baseline and every 2 years during the following 12 years.

For the present study, a cross-sectional analysis was performed with data from the 8-year visit (n=135), which is the first time point where all the outcome measurement instruments of interest were included. For the descriptive analysis, patients were included if they had data of at least one of the measurement instruments of interest for each domain, and for the construct validity analysis, only those with complete data from all the instruments for each domain as well as data from all the external constructs were included.

Pain and morning stiffness domains and external constructs

In OASIS, three self-reported single items were available to assess the domain back pain: (1) total back pain (‘*How much back pain did you have during the past week?*’), (2) back pain at night (‘*How much back pain did you have at night during the past week?*’) and (3) total back pain by BASDAI-Q2[2] (‘*How would you describe the overall level of AS neck back or hip you have had during the last week?*’). Three items from BASDAI[2] were used for the morning stiffness domain: (1) morning stiffness-severity (BASDAI-Q5; ‘*How would you describe the level of* morning stiffness *you have had from the time you wake up during the last week?*’), (2) morning stiffness-duration (BASDAI-Q6; ‘*How long did your* morning stiffness *last from the time you wake up during the last week?*’) and (3) morning stiffness-severity/duration (BASDAI-Q5/6). Each item was scored on a 0–10 NRS, with 0 representing absence of pain/morning stiffness and 10 very severe pain/morning stiffness.

The following outcomes were used as external constructs in the construct validity analysis of each of the pain and morning stiffness instruments: BASDAI, Axial Spondyloarthritis Disease Activity Score (ASDAS),[7] C reactive protein (mg/L) and erythrocyte sedimentation rate (mm/hour), swollen joint count and tender joint count, fatigue, Patient Global Assessment (PGA), Physician Global Assessment (PhGA), Bath Ankylosing Spondylitis Global Score, morning stiffness (BASDAI-Q5/6) and fatigue (BASDAI-Q1) for disease activity; the Bath Ankylosing Spondylitis Functional Index (BASFI)[11] for functional ability; Bath Ankylosing Spondylitis Mobility Index (BASMI)[12] for spinal mobility; the rating scale of the EuroQoL-5 dimensions,[13] Ankylosing Spondylitis Quality of Life,[14] Short Form 36 Physical Component Summary and Mental Component Summary[15] for quality of life and modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS)[16] for structural damage. The hypotheses for the association between each of external constructs and the measurement instruments of interest were extracted from the ASAS-COS exercise. For total back pain, the hypotheses from total back pain BASDAI-Q2 were assumed.[1]

To assess known group discrimination, several subgroups of patients were created by dichotomising the external constructs by the established cut-offs (ASDAS<2.1 vs ≥2.1; BASDAI<4 vs ≥4) or by the median value (PGA<4 vs ≥4; PhGA<4 vs ≥4, fatigue<5 vs ≥5, low vs high functional impairment (BASFI<4 vs ≥4), low vs high spinal mobility impairment (BASMI<4 vs ≥4) and low vs high spinal structural damage (mSASSS<15 vs ≥15)).

Potential contextual factors

Sociodemographic data (including age, sex, educational level (low: no education, primary school or lower secondary school; middle: upper secondary school or post-secondary non-tertiary school; and high: bachelor’s, tertiary, master’s or doctoral education), body mass index (BMI)) as well as disease characteristics (symptom duration, type of axSpA involvement—pure axial or axial and peripheral) were also collected. Each of them was dichotomised by the median value or by the investigators’ discretion based on common practice: age (dichotomised using the WHO definition (<60 vs ≥60)), sex (male vs female), educational level (high education vs other) and BMI (global cut-off for normal/overweight (<25 vs ≥25 kg/m^2)).

Statistical analysis

The distribution of scores was examined for each instrument of interest, as well as the frequency of missing data and the identification of floor and ceiling effects. The construct validity of the instruments was analysed through Spearman correlations between the pain/morning stiffness instruments and other relevant disease domains. The correlation cut-offs used were weak <0.30, moderate 0.30–0.69 and strong ≥0.70.[17] Taking the hypotheses on the strength of the correlation from the axSpA COS,[1] we could assess (independently for each instrument) whether they fulfilled the threshold of 75% of the hypotheses confirmed (ie, good performance), used for the ASAS-COS.

Additionally, as part of the construct validity assessment, known group discrimination was analysed in order to compare and identify which of the instruments discriminate the best between groups dichotomised based on the external constructs as described above. The discriminatory capacity of each pain and morning stiffness instrument was assessed by calculating the standardised mean difference (SMD), which corresponds to the difference of the group means divided by the pooled SD of the group means. It is unitless, with a higher absolute value reflecting a higher discriminatory capacity (ie, distinguishing between groups). An SMD≥0.80 corresponds to a large discriminatory capacity, ≥0.50 moderate discriminatory capacity and >0.2 small discriminatory capacity at best.[18]

In order to assess the potential influence of contextual factors on known group discrimination of the pain and morning stiffness instruments, with the focus on the ones with the highest discriminatory capacity, all SMDs were recalculated in subgroups of patients dichotomised based on the contextual factors. This allowed to assess the ‘measurement affecting’ property, that is, whether the known group discrimination of the measurement instrument is affected by different contextual factors. In order to understand whether the difference between the SMDs from the two subgroups was significant, the 95% CI of each contrast or difference was calculated.[19] All the analyses were performed using Stata SE V.17.

RESULTS

Patient characteristics

Of 135 patients with 8-year follow-up in the OASIS cohort, 98 (73%) had data available on at least one of the instruments of interest for each domain and could be included. Of these, 71% were male, with a mean age

of 54 years (SD 11) and a mean symptom duration (8 years after the inclusion in the cohort) of 31 (11) years (table 1). The mean values for total back pain, back pain at night and total back pain BASDAI-Q2 were 3.7 (2.3), 2.9 (2.3) and 4.6 (2.6), respectively. As for morning stiffness, the mean values of morning stiffness-severity (BASDAI-Q5), morning stiffness-duration (BASDAI-Q6) and morning stiffness-severity/duration (BASDAI-Q5/6) were 3.7 (2.6), 3.3 (3.1) and 3.5 (2.7), respectively.

Interpretability, floor and ceiling effect

The distribution of score is visualised in online supplemental figure S1. There were three patients with missing data for the question of back pain at night, and one patient missing all the morning stiffness questions (and therefore excluded). For the pain domain, the ceiling effects were negligible (≤1%). Floor effect was highest for back pain at night (15%; 14/98), followed by total back pain (10%; 10/98) and total back pain BASDAI-Q2 (4%; 4/98). For the morning stiffness domain, ceiling effects for morning stiffness-severity and morning stiffness-severity/duration were small (1%; 1/98) but 8%; 8/98 for the morning stiffness-duration. Floor effect was largest for morning stiffness-duration (12%; 12/98), followed by morning stiffness-severity/duration and morning stiffness-severity (10%; 10/98).

Construct validity

Construct validity was analysed in 85 patients with complete data for the six instruments as well as external constructs. Sociodemographic and disease characteristics did not differ from the total group (table 1). The correlations between pain and morning stiffness instruments and the different external constructs (health outcomes) are presented in table 2, with the hypotheses of correlation strength previously established that were confirmed in bold. For pain instruments, only back pain at night had 75% of the hypotheses confirmed (good performance). Although total back pain BASDAI-Q2 and back pain reached only 62% and 69%, respectively (adequate performance), when accepting that correlations that were higher than expected also meet the hypotheses, these instruments also passed the threshold of ≥75% confirmations. Similarly, all the morning stiffness instruments passed from adequate to good performance when considering a better correlation than hypothesised as confirmation.

Known group discrimination

The discriminatory capacity between disease activity states, defined by the external constructs ASDAS and BASDAI, was large for all pain instruments (SMD≥0.80) although higher for total back pain BASDAI-Q2 and total back pain compared with back pain at night, with total back pain BASDAI-Q2 performing mostly slightly better: SMD 1.87 vs 1.56 vs 1.07 when categorised by ASDAS (figure 1), and 2.31 vs 1.83 vs 1.32 by BASDAI. Tables 3 and 4 present, for each of the pain and morning stiffness

Table 1 Patient characteristics at 8-year visit

	n=98*	n=85†
Age (years)	54 (11)	54 (11)
Male sex	69 (71%)	60 (71%)
Symptom duration (years)	31.2 (11.3)	31.4 (11.0)
Disease duration (years)‡	20.5 (9.4)	20.3 (9.2)
HLA-B27 positive‡	81 (86%)	71 (86%)
Pure axial disease	73 (82%)	70 (82%)
Back pain (0–10)	3.7 (2.3)	3.7 (2.3)
Back pain at night (0–10)	2.9 (2.3)	3.1 (2.3)
Total back pain (BASDAI-Q2) (0–10)	4.6 (2.6)	4.7 (2.6)
Morning stiffness-severity (BASDAI-Q5) (0–10)	3.7 (2.6)	3.7 (2.5)
Morning stiffness-duration (BASDAI-Q6) (0–10)	3.3 (3.1)	3.3 (3.1)
Morning stiffness-severity/duration (BASDAI-Q5/6) (0–10)	3.5 (2.7)	3.5 (2.7)
BASDAI (0–10)	3.8 (2.2)	3.8 (2.1)
ASDAS	2.5 (1.0)	2.4 (1.0)
CRP (mg/L)	8.0 (8.4)	8.1 (8.4)
ESR, mm/hour‡	17.8 (16.4)	17.9 (16.4)
Fatigue (BASDAI-Q1) (0–10)	5.0 (2.6)	5.1 (2.5)
PGA (0–10)	4.0 (2.6)	4.1 (2.5)
PhGA (0–10)	3.8 (2.5)	3.8 (2.5)
BAS-G (0–10)	4.1 (2.4)	4.2 (2.3)
BASFI (0–10)	4.2 (2.5)	4.2 (2.4)
BASMI (0–10)	4.1 (1.6)	4.1 (1.6)
EuroQoL thermometer (0–100)‡	64 (20)	64.4 (20.0)
ASQoL (0–18)‡	6.7 (4.8)	6.8 (4.6)
SF-36 PCS	39.2 (11.5)	39.1 (11.4)
SF-36 MCS	49.0 (12.4)	48.6 (12.2)
Employed (≤65 years)‡	36 (44%)	32 (45%)
Blue collar (≤65 years)§	37 (58%)	31 (56%)
Educational level‡		
Low education	41 (43%)	37 (44%)
Middle education	37 (38%)	32 (38%)
High education	18 (19%)	15 (18%)
BMI (kg/m^2)‡	26.1 (4.4)	26.2 (4.4)
mSASSS (0–72)§	20.7 (19.8)	21.2 (19.9)
NSAID, n (%)‡	67 (72)	61 (72)
Biologics, n (%)	19 (20)	17 (20)

Data are presented as mean (SD) or n (%).
*Patients with data available on at least one of the instruments of interest for each domain.
†Patients with complete data for the 6 instruments and external constructs.
‡<5% of the data are missing.
§<25% of missing data.
ASDAS, Axial Spondyloarthritis Disease Activity Score; ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; BAS-G, Bath Ankylosing Spondylitis Global Score; BASMI, Bath Ankylosing Spondylitis Metrology Index; BMI, body mass index; CRP, C reactive protein; ESR, erythrocyte sedimentation rate; EuroQoL, Euro Quality of Life; HLA, human leukocyte antigen B27; HLA-B27, Human Leukocyte Antigen B27; mSASSS, modified Stoke Ankylosing Spondylitis Spinal Score; NSAID, non-steroidal anti-inflammatory drug; PGA, Patient Global Assessment; PhGA, Physician Global Assessment; SF-36 PCS and MCS, Short Form 36 Physical Component Summary and Mental Component Summary.

Table 2 Spearman correlations between the pain and morning stiffness measures and the different external constructs (n=85)

External comparators	Strength of correlation and alignment with a priori hypothesised correlation strength (Weak – Moderate – Strong)					
	Total back pain	Back pain at night	Total back pain BASDAI-Q2	Morning stiffness severity (BASDAI-Q5)	Morning stiffness duration (BASDAI-Q6)	Morning stiffness severity/ duration (BASDAI-Q5/6)
BASDAI	**Strong**	**Strong**	**Strong**	Moderate	Moderate	Moderate
	0.786	0.701	0.848	0.819	0.738	0.795
ASDAS	Strong	Strong	**Strong**	**Moderate**	**Moderate**	**Moderate**
	0.676	0.545	0.773	0.645	0.598	0.635
CRP, (mg/l)	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**
	0.037	−0.016	0.099	−0.011	0.012	−0.001
ESR, mm/h	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**
	−0.094	−0.025	−0.050	−0.116	−0.113	−0.109
SJC	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**
	0.139	0.142	0.190	0.246	0.151	0.188
TJC						
	0.417	0.392	0.438	0.412	0.416	0.429
Total back pain				Strong	Strong	Strong
				0.691	0.572	0.644
PGA	**Strong**	**Moderate**	Strong	Moderate	**Moderate**	**Moderate**
	0.757	0.634	0.692	0.724	0.554	0.664
PhGA	**Moderate**	**Moderate**	**Moderate**	Weak	Weak	Weak
	0.699	0.638	0.674	0.712	0.559	0.652
Morning stiffness	Strong	**Moderate**	Strong			
	0.644	0.631	0.659			
Fatigue	**Moderate**	**Moderate**	Moderate	**Moderate**	**Moderate**	**Moderate**
	0.670	0.561	0.702	0.679	0.545	0.621
BAS-G	Moderate	**Moderate**	Moderate	Moderate	**Moderate**	**Moderate**
	0.784	0.697	0.740	0.736	0.600	0.684
BASFI	**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate**
	0.487	0.469	0.488	0.517	0.476	0.507
BASMI	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
	−0.075	−0.077	0.006	−0.005	−0.093	−0.044
ASQoL	**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate**
	0.520	0.533	0.544	0.601	0.531	0.578
SF-36 PCS	Moderate	Moderate	**Moderate**	Weak	**Weak**	Weak
	−0.419	−0.392	−0.423	−0.401	−0.291	−0.357
SF-36 MCS	**Weak**	Weak	Weak	Weak	Weak	Weak
	−0.211	−0.399	−0.422	−0.416	−0.339	−0.390
mSASSS	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**	**Weak**
	−0.160	−0.194	−0.036	−0.068	−0.055	−0.063
Construct validity*	11/16=69%	12/16=75%	10/16=62%	8/16=50%	11/16=69%	10/16=64%
Construct validity if correlation is higher than hypothesised	12/16=75%	13/16=81%	13/16=81%	14/16=88%	14/16=88%	14/16=88%

Weak <0.30. Moderate 0.30–0.69. Strong ≥0.70. Hypotheses extracted from the instrument selection for the ASAS Core Outcome Set for axial spondyloarthritis.[1] Bold denotes the hypothesis was confirmed; white cell when there was no hypothesis.
*Construct validity: ≥75% of the hypotheses confirmed: good (green); 50–75% of the hypotheses confirmed: adequate (orange); <50% of the hypotheses confirmed: poor (red).
ASAS, Assessment of SpondyloArthritis international Society; ASDAS, Axial Spondyloarthritis Disease Activity Score; ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; BAS-G, Bath Ankylosing Spondylitis Global Score; BASMI, Bath Ankylosing Spondylitis Metrology Index; CRP, C reactive protein; ESR, erythrocyte sedimentation rate; mSASSS, modified Stoke Ankylosing Spondylitis Spinal Score; PGA, Patient Global Assessment; PhGA, Physician Global Assessment; SF-36 PCS and MCS, Short Form 36 Physical Component Summary and Mental Component Summary; SJC, swollen joint count; TJC, tender joint count.

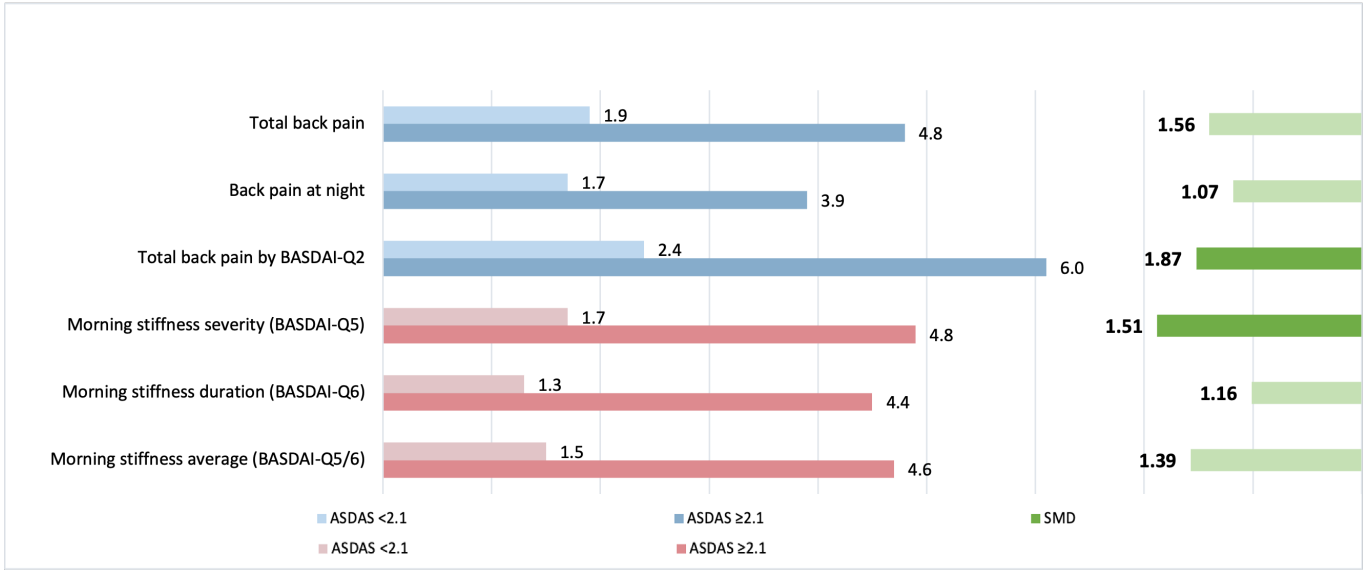

Figure 1 Discrimination of pain and morning stiffness measurement instruments between low and high disease activity by Axial Spondyloarthritis Disease Activity Score (ASDAS). Blue columns (pain domain) and pink columns (morning stiffness domain) on the left side represent the mean value in the low (light blue) and high (dark blue) disease activity strata. Columns in green on the right side depict the standardised mean difference (SMD) to discriminate between the strata for each measurement instrument. The higher the SMD, the higher the discriminatory capacity. BASDAI-Q2, Bath Ankylosing Spondylitis Disease Activity Index-Question 2.

instruments, the results of the mean and SD within each of the subgroups defined by dichotomised external constructs, as well as the SMDs reflecting the discrimination of each pain/morning stiffness instrument between subgroups. For example, the mean (SD) of total back pain BASDAI-Q2 in patients with ASDAS<2.1 was 2.45 (1.67), whereas in patients with an ASDAS≥2.1 the mean was 6.02 (2.03), resulting in an SMD of 1.87. Total back pain BASDAI-Q2 was also the measure with the highest SMD, thus the largest discriminatory capacity, between groups stratified based on fatigue, functional ability and spinal mobility. As an exception, total back pain best discriminated between health states by PGA and PhGA. It should be noted that none of the pain instruments discriminated well between subgroups that differed in spinal mobility and spinal radiographic damage.

Regarding morning stiffness, morning stiffness-severity compared with morning stiffness-severity/duration and morning stiffness-duration had consistently higher SMDs across all health states of external constructs, except for BASDAI, where the morning stiffness-severity/duration performed slightly better. Once again, the discriminatory capacity was large for all the external constructs except for spinal mobility and spinal radiographic damage. On the other hand, morning stiffness-duration consistently performed the worst across all the external constructs.

Contextual factors

When considering the instruments with the best known group discrimination (total back pain BASDAI-Q2 and morning stiffness-severity), limited influences of contextual factors on SMDs were observed. For *age*, no other significant influence was found on the discriminatory capacity of any of instrument for pain or morning stiffness-severity across different external constructs. For example, the SMD for total back pain by BASDAI-Q2 in persons <60 years old between active and non-active disease based on ASDAS was 1.94, while in those ≥60 years old was 1.72. The contrast *or difference* between them was 0.22 (95% CI −0.86, 1.30), meaning that even though the discriminatory capacity of total spinal pain by BASDAI-Q2 to discriminate between the two ASDAS states was better for younger individuals, the difference (vs older individuals) was not statistically significant (online supplemental table S2). For age, also stratifying by *sex* could not reveal significant differences in known group discrimination (online supplemental table S3). For *education,* however, the SMD of total back pain by BASDAI-Q2 between high and low PGA indicated was very large (SMD 2.05) in persons non-highly educated and small (SMD 0.47) for high-educated persons (contrast: −1.58 (95% CI −2.76, −0.40)). Similar results were found for PhGA, although not statistically significant (SMD 0.32 vs 1.45; contrast: −1.13 (95% CI −2.30, 0.04)) (online supplemental table S4). Lastly, *BMI* significantly influenced the SMDs for morning stiffness-severity and morning stiffness-severity/duration, when discriminating disease activity (by ASDAS, BASDAI and PGA (only morning stiffness-severity)). Of note, although all these instruments had a large discriminatory capacity in both strata (overweight and not-overweight), there were statistically significant differences in morning stiffness-severity and morning stiffness-severity/duration, with higher discriminatory capacity in the not-overweight versus the overweight population (online supplemental table S5).

Table 3 Discrimination of pain and morning stiffness measurement instruments between patients with low versus high disease activity

Assessment measure	ASDAS<2.1 (n=31)	ASDAS≥2.1 (n=54)	SMD	BASDAI<4 (n=47)	BASDAI≥4 (n=38)	SMD	PGA<4 (n=38) (SD)	PGA≥4 (n=47) (SD)	SMD	Low PhGA (<4) (n=43)	High PhGA (≥4) (n=42)	SMD
Total back pain	1.94 (1.59)	4.78 (1.95)	1.56	2.36 (1.67)	5.45 (1.70)	1.83	2.05 (1.47)	5.11 (1.87)	**1.79**	2.46 (1.64)	5.05 (2.11)	**1.37**
Back pain at night	1.68 (1.74)	3.87 (2.20)	1.07	1.94 (1.59)	4.47 (2.26)	1.32	1.71 (1.50)	4.17 (2.24)	1.26	1.91 (1.58)	4.26 (2.31)	1.19
Total back pain (BASDAI-Q2)	2.45 (1.67)	6.02 (2.03)	**1.87**	2.98 (1.84)	6.87 (1.47)	**2.31**	2.97 (2.07)	6.13 (2.01)	1.55	3.44 (2.25)	6.02 (2.20)	1.16
Morning stiffness-severity (BASDAI-Q5)	1.74 (1.46)	4.81 (2.29)	**1.51**	2.08 (1.41)	5.68 (2.11)	2.05	1.84 (1.30)	5.19 (2.23)	**1.78**	2.09 (1.48)	5.33 (2.28)	**1.69**
Morning stiffness-duration (BASDAI-Q6)	1.29 (1.24)	4.44 (3.26)	1.16	1.38 (1.24)	5.66 (3.08)	1.90	1.63 (1.63)	4.64 (3.35)	1.10	1.77 (1.70)	4.88 (3.42)	1.15
Morning stiffness-severity/duration (BASDAI-Q5/6)	1.52 (1.26)	4.63 (2.63)	1.39	1.73 (1.19)	5.67 (2.41)	**2.14**	1.74 (1.38)	4.91 (2.65)	1.46	1.93 (1.50)	5.10 (2.70)	1.46

Values shown as mean (SD). Best performing SMD shown in bold.
ASDAS, Axial Spondyloarthritis Disease Activity Score; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; PGA, Patient Global Assessment; PhGA, Physician Global Assessment; SMD, standardised mean difference.

Table 4 Discrimination of pain and morning stiffness measurement instruments between patients with low versus high fatigue, functional ability, spinal mobility impairment and spinal radiographic damage

Assessment measure	Fatigue<5 (n=35)	Fatigue≥5 (n=50)	SMD	BASFI<4 (n=38)	BASFI≥4 (n=47)	SMD	BASMI<4 (n=39) Mean (SD)	BASMI≥4 (n=46) Mean (SD)	SMD	mSASSS<15 (n=31) Mean (SD)	mSASSS≥15 (n=34) Mean (SD)	SMD
Total back pain	2.11 (1.69)	4.88 (1.92)	1.51	2.74 (2.06)	4.55 (2.13)	0.86	3.51 (2.21)	3.93 (2.34)	0.18	3.9 (2.3)	2.2 (2.3)	**−0.32**
Back pain at night	1.66 (1.39)	4.06 (2.29)	1.22	2.13 (1.96)	3.83 (2.28)	0.79	3.08 (2.28)	3.06 (2.33)	−0.005	3.0 (2.1)	2.6 (2.3)	−0.20
Total back pain (BASDAI-Q2)	2.86 (1.97)	6.02 (2.09)	**1.55**	3.58 (2.62)	5.64 (2.14)	**0.87**	4.44 (2.59)	4.96 (2.55)	**0.20**	5.0 (2.7)	4.2 (2.5)	−0.29
Morning stiffness-severity (BASDAI-Q5)	1.91 (1.44)	4.94 (2.34)	**1.50**	2.50 (2.08)	4.66 (2.42)	**0.95**	3.43 (2.36)	3.91 (2.63)	**0.19**	3.8 (2.6)	3.6 (2.6)	−0.11
Morning stiffness-duration (BASDAI-Q6)	1.63 (1.65)	4.46 (3.34)	1.02	2.00 (2.20)	4.34 (3.33)	0.81	3.20 (3.06)	3.37 (3.16)	0.05	3.4 (3.4)	3.0 (2.4)	**−0.14**
Morning stiffness-severity/duration (BASDAI-Q5/6)	1.77 (1.45)	4.70 (2.70)	1.29	2.25 (2.04)	4.50 (2.74)	0.92	3.32 (2.58)	3.64 (2.79)	0.12	3.5 (2.9)	3.3 (2.6)	−0.13

Values shown as mean (SD). Best performing SMD shown in bold.
BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; BASMI, Bath Ankylosing Spondylitis Metrology Index; mSASSS, modified Stoke Ankylosing Spondylitis Spinal Score; SMD, standardised mean difference.

DISCUSSION

This comprehensive analysis comparing the construct validity across instruments for pain and morning stiffness in patients with axSpA is supportive of the choices made in the axSpA COS. When accepting that a better correlation than hypothesised is actually a positive finding, the total back pain BASDAI-Q2 has good construct validity. Moreover, total back pain BASDAI-Q2 had more favourable SMD and a lower floor effect over back pain at night and back pain, turning it into the preferred pain instrument. For morning stiffness, even though the morning stiffness-severity (BASDAI-Q5) was slightly superior, morning stiffness-severity/duration (BASDAI-Q5/6) had a very good performance as well and can be used in order to increase the information with two aspects of the symptom (severity and duration). These instruments were mostly stable across contextual factor, with the only exceptions being the influence of educational level on total back pain BASDAI-Q2 when discriminating between high and low PGA in high-educated individuals, and BMI on morning stiffness-severity and morning stiffness-severity/average.

During the update of the axSpA COS, domain match, feasibility, construct validity and discrimination (referring to reliability, responsiveness and discrimination in clinical trials) were assessed. However, the construct validity assessment was only based on correlations to test hypotheses for the strength of the correlation between instruments. Although this approach is very informative, it has limitations as it remains subjective with large variation on hypotheses across experts, besides the lack of clinical relevance. This becomes especially challenging when there are several instruments for the same domain and the aim is to recommend one preferred instrument per domain, as it is the case of the ASAS-COS. Known group discrimination can be a useful method when deciding on the best measurement instrument, allowing a fair comparison and selection of the instruments. Furthermore, this is typically being assessed when a measurement instrument is developed, like ASDAS,[7 20] the Assessment of SpondyloArthritis international Society Health Index (ASAS-HI)[21 22] in axSpA or the Disease Activity in PSoriatic Arthritis with 28-joint count (DAPSA28) in psoriatic arthritis,[23] since such analysis gives more insight into the real meaning of the values of the newly developed score.

Our analysis of construct validity by predefined hypotheses of correlations showed that, in line with the COS and strictly speaking, back pain at night had a good performance, with >75% of the hypotheses confirmed, while total back pain BASDAI-Q2, along with the three morning stiffness instruments, had adequate performance. Of note, most of the 'rejected' hypotheses in these instruments were due to a higher correlation than what was hypothesised, and therefore, when accepting these scenarios as hypotheses met, all six instruments showed a good performance. However, a main issue is that the hypothesis we used had been determined by the ASAS-COS initiative and was formulated for

each individual instrument and not for the purpose of comparison. This could partly explain why some hypotheses were surprisingly different across instruments of the same domain. This supports the approach to include a quantitative approach when comparing validity of instruments addressing the same domain and include known group discrimination.

The effect of some contextual factors on health outcomes has been previously reported. For example, female sex and low education have been associated with higher disease activity, worse functional ability and worse overall functioning and health (outcome influencing).[24 25] However, to the best of our knowledge, notwithstanding recommended by OMERACT,[8] there are no data assessing the impact of those contextual factors on the psychometric properties of the instruments (measurement affecting). When moving to 'stratified' medicine, it is important to understand whether subgroup differences can be explained by measurement issues (contextual factors affecting the external construct) or have a true influence on the outcome of interest. However, there is no standardised or formal recommendation on how to assess such potential effects. Thereby, we decided to replicate the known group discrimination analysis in different contextual factor subgroups and to formally compare the differences in terms of statistical significance of the magnitude. Educational level was the most striking contextual factor that had a significant effect and showed that total back pain BASDAI-Q2 had better discriminatory capacity in non-highly educated individuals than in those high educated. On a same line, BMI influenced known group discrimination of morning stiffness-severity and morning stiffness-severity/average for disease activity stated by BASDAI. It should be kept in mind that the numbers in each subgroup were very small. Recently, the effect of contextual factors on the measurement properties was reported in a study assessing the effect of contextual factors on performance of thresholds for presenteeism instruments.[26] We recommend to further test the role of contextual factor on measurement properties of existing or new instruments in sufficiently powered studies. Our approach can serve as an example, but we should realise that this requires large populations.

In terms of study limitations, besides the small sample size leading to small number of patients for the contextual factor analysis, another limitation of the study is that it was performed in a cohort of patients with established disease with a long disease duration, representing only patients with r-axSpA. It has already been demonstrated that the burden of the disease is comparable between r-axSpA and non-r-axSpA,[27] as so is the performance of several instruments[1]; therefore, the same construct validity of the assessed instruments would be expected. Nevertheless, to ensure generalisability of the findings, a similar analysis should be conducted in earlier phases of the disease like in early axSpA (≤2 years of axial symptoms).[28] Being the main aim of this analysis to compare the performance of different instruments, one may wonder to what extent the hypotheses for the correlations should have been derived taking this into account. On the other hand, we decided to keep the same hypotheses as used for the axSpA COS to not introduce more confusion to this type of 'abstract' assessment. Another potential limitation was that the analysis was performed using fully available data from the 8-year visit instead of baseline data, which could introduce a selection bias. However, no major differences were found when comparing baseline characteristics from various samples (ie, full baseline data, full 8-year visit data, etc) (online supplemental table S1), thus reducing the risk of bias.

In conclusion, we confirmed construct validity for the recommended ASAS-COS pain instrument total back pain BASDAI-Q2 and morning stiffness instrument morning stiffness-severity/duration. Back pain BASDAI-Q2 was confirmed to discriminate best between patients with high and low disease activities, as well as other relevant disease domains. Morning stiffness-severity/duration performed well and this was also the case for morning stiffness-severity, with an even slightly better performance. Educational level may influence the total back pain BASDAI-Q2 performance as well as the BMI may influence the performance of the morning stiffness instruments, but otherwise contextual factor did not modify the performance of pain or morning stiffness instruments.

Author affiliations
[1]Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, Netherlands
[2]Rheumatology, Tel Aviv Sourasky Medical Center–Ichilov, Tel Aviv, Israel
[3]Centre for Rheumatic Diseases, King's College Hospital Charity, London, UK
[4]Rheumatology, King's College Hospital Charity, London, UK
[5]Department of Rheumatology, Maastricht University Medical Centre+, Maastricht, Netherlands
[6]Section for Biostatistics and Evidence-Based Research, Bispebjerg Hospital, Kobenhavn, Denmark
[7]Clinical Research, Odense University Hospital, Odense, Denmark
[8]Rheumatology, Leiden University Medical Center, Leiden, Netherlands
[9]Clinical Immunology and Rheumatology, Amsterdam University Medical Centres, Amsterdam, Netherlands
[10]Rheumatology, Zuyderland Medical Centre, Sittard-Geleen, Netherlands
[11]Department of Rheumatology, Zuyderland Medical Centre Heerlen, Heerlen, Netherlands

Contributors SR, AB, EN and DC designed the study. SR, AB, AvT, RL and DvdH collected and/or prepared the data. DC and RC analysed the data. DC, SR, AB, EN and RC critically interpreted the results. DC, SR, AB and EN were involved in drafting the manuscript. All authors revised the manuscript critically for important intellectual content and approved the final manuscript. DC is responsible for the overall content as the guarantor.

Funding RC (Section for Biostatistics and Evidence-Based Research, Parker Institute at Bispebjerg and Frederiksberg Hospital) is supported by a core grant from the Oak Foundation (OCAY-18-774-OFIL).

Competing interests EN received research grants from AbbVie, Alfasigma, Fresenius, Pfizer, and UCB, and honoraria or consultancy/speaker fees from AbbVie, Eli Lilly, Alfasigma, Pfizer and UCB. AB received research grants from AbbVie and Eli Lilly and honoraria or consultancy/speaker fees from Pfizer, AbbVie, Galapagos, UCB and Novartis. DvdH received consultant fees from AbbVie, Argenx, BMS, Galapagos, GlaxoSmithKline, Janssen, Lilly, Novartis, Pfizer, Takeda and UCB Pharma. RL received consultant fees from AbbVie, BMS, Galapagos, Eli Lilly, Novartis, Janssen, Pfizer and UCB. AvT received research grants and consultant fees from Novartis and Galapagos. SR received research grants from AbbVie, Galapagos, MSD, Novartis, Pfizer, and UCB, and honoraria or consultancy/speaker

fees from AbbVie, Eli Lilly, Galapagos, Janssen, MSD, Novartis, Pfizer, UCB and Sanofi.

Patient consent for publication Not applicable.

Ethics approval The ethics committees from all participating hospitals have approved the study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Dafne Capelusnik http://orcid.org/0000-0001-9336-0416
Elena Nikiphorou http://orcid.org/0000-0001-6847-3726
Annelies Boonen http://orcid.org/0000-0003-0682-9533
Désirée van der Heijde http://orcid.org/0000-0002-5781-158X
Robert Landewé http://orcid.org/0000-0002-0577-6620
Astrid van Tubergen http://orcid.org/0000-0001-8477-0683
Sofia Ramiro http://orcid.org/0000-0002-8899-9087

REFERENCES

1 Navarro-Compán V, Boel A, Boonen A, *et al*. Instrument selection for the ASAS core outcome set for axial spondyloarthritis. *Ann Rheum Dis* 2023;82:763–72.
2 Garrett S, Jenkinson T, Kennedy LG, *et al*. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994;21:2286–91.
3 Sieper J, Rudwaleit M, Baraliakos X, *et al*. The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. *Ann Rheum Dis* 2009;68:ii1–44.
4 Maxwell LJ, Beaton DE, Boers M, *et al*. The evolution of instrument selection for inclusion in core outcome sets at OMERACT: Filter 2.2. *Semin Arthritis Rheum* 2021;51:1320–30.
5 Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
6 Leung YY, Orbai AM, Hojgaard P, *et al*. OMERACT Filter 2.1 instrument selection for physical function domain in psoriatic arthritis: Provisional endorsement for HAQ-DI and SF-36 PF. *Semin Arthritis Rheum* 2021;51:1117–24.
7 van der Heijde D, Lie E, Kvien TK, *et al*. ASDAS, a highly discriminatory ASAS-endorsed disease activity score in patients with ankylosing spondylitis. *Ann Rheum Dis* 2009;68:1811–8.
8 Nielsen SM, Boers M, de Wit M, *et al*. OMERACT consensus-based operational definition of contextual factors in rheumatology clinical trials: A mixed methods study. *Semin Arthritis Rheum* 2021;51:601–6.
9 Nielsen SM, Uggen Rasmussen M, Boers M, *et al*. Towards consensus in defining and handling contextual factors within rheumatology trials: an initial qualitative study from an OMERACT working group. *Ann Rheum Dis* 2021;80:242–9.
10 Spoorenberg A, van der Heijde D, de Klerk E, *et al*. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol* 1999;26:980–4.
11 Calin A, Garrett S, Whitelock H, *et al*. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994;21:2281–5.
12 Jenkinson TR, Mallorie PA, Whitelock HC, *et al*. Defining spinal mobility in ankylosing spondylitis (AS). *Bath AS Metrology Index J Rheumatol* 1994;21:1694–8.
13 EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199–208.
14 Doward LC, Spoorenberg A, Cook SA, *et al*. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Ann Rheum Dis* 2003;62:20–6.
15 Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993;306:1437–40.
16 van der Heijde D, Braun J, Deodhar A, *et al*. Modified stoke ankylosing spondylitis spinal score as an outcome measure to assess the impact of treatment on structural progression in ankylosing spondylitis. *Rheumatology (Oxford)* 2019;58:388–400.
17 Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012;24:69–71.
18 Murad MH, Wang Z, Chu H, *et al*. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ* 2019;364:k4817.
19 Christensen R, Bours MJL, Nielsen SM. Effect Modifiers and Statistical Tests for Interaction in Randomized Trials. *J Clin Epidemiol* 2021;134:174–7.
20 Lukas C, Landewé R, Sieper J, *et al*. Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. *Ann Rheum Dis* 2009;68:18–24.
21 Kiltz U, van der Heijde D, Boonen A, *et al*. The ASAS Health Index (ASAS HI) - a new tool to assess the health status of patients with spondyloarthritis. *Clin Exp Rheumatol* 2014;32:S–105.
22 Kiltz U, van der Heijde D, Boonen A, *et al*. Clinimetric Validation of the Assessment of Spondyloarthritis International Society Health Index in Patients With Radiographic Axial Spondyloarthritis in Ixekizumab Trials. *J Rheumatol* 2023;50:754–62.
23 Michelsen B, Sexton J, Smolen JS, *et al*. Can disease activity in patients with psoriatic arthritis be adequately assessed by a modified Disease Activity index for PSoriatic Arthritis (DAPSA) based on 28 joints? *Ann Rheum Dis* 2018;77:1736–41.
24 Capelusnik D, Zhao SS, Boonen A, *et al*. Individual-level and country-level socio-economic factors and health outcomes in spondyloarthritis: analysis of the ASAS-perSpA study. *Rheumatology (Oxford)* 2022;61:2043–53.
25 Putrik P, Ramiro S, Moltó A, *et al*. Individual-level and country-level socioeconomic determinants of disease outcomes in SpA: multinational, cross-sectional study (ASAS-COMOSPA). *Ann Rheum Dis* 2019;78:486–93.
26 Capelusnik D, Ramiro S, Nikiphorou E, *et al*. Thresholds for unacceptable work state in radiographic axial Spondyloarthritis of four presenteeism and two clinical outcome measurement instruments. *Rheumatology (Oxford)* 2024.:keae033.
27 López-Medina C, Ramiro S, van der Heijde D, *et al*. Characteristics and burden of disease in patients with radiographic and non-radiographic axial Spondyloarthritis: a comparison by systematic literature review and meta-analysis. *RMD Open* 2019;5:e001108.
28 Navarro-Compán V, Benavent D, Capelusnik D, *et al*. ASAS consensus definition of early axial spondyloarthritis. *Ann Rheum Dis* 2023;ard–2023.