RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Test–retest reliability of outcome measures: data from three trials in radiographic and non-radiographic axial spondyloarthritis

Anne Boel [1], Victoria Navarro-Compán,[2] Désirée van der Heijde [1]

To cite: Boel A, Navarro-Compán V, van der Heijde D. Test–retest reliability of outcome measures: data from three trials in radiographic and non-radiographic axial spondyloarthritis. *RMD Open* 2021;**7**:e001839. doi:10.1136/rmdopen-2021-001839

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2021-001839).

Received 22 July 2021
Accepted 4 November 2021

[1]Rheumatology Department, Leiden University Medical Center, Leiden, The Netherlands
[2]Rheumatology Department, La Paz University Hospital, Madrid, Spain

Correspondence to
Ms Anne Boel;
a.h.e.m.boel@lumc.nl

ABSTRACT

Objectives Aim of this study was to assess test–retest reliability of candidate instruments for the mandatory domains of the Assessment of Spondyloarthritis international Society (ASAS)-Outcome Measures in Rheumatology core set for axial spondyloarthritis (axSpA).

Methods Screening and baseline data from COAST-V, COAST-X and RAPID-axSpA was used to evaluate test–retest reliability of each candidate instrument for the mandatory domains (disease activity, pain, morning stiffness, fatigue, physical function, overall functioning and health). A maximum time interval of 28 days between both visits was used for inclusion in this study. Test–retest reliability was assessed by intraclass correlation coefficient (ICC). Bland and Altman plots provided mean difference and 95% limits of agreement, which were used to calculate the smallest detectable change (SDC). Data were analysed for radiographic and non-radiographic axSpA separately.

Results Good reliability was found for Ankylosing Spondylitis Disease Activity Score (ICC 0.79, SDC 0.6), C reactive protein (ICC 0.72–0.79, SDC 12.3–17.0), Bath Ankylosing Spondylitis Functional Index (ICC 0.87, SDC 1.1) and 36-item Short-Form Health Survey (ICC Physical Component Summary 0.81, SDC 4.7, Mental Component Summary 0.80, SDC 7.3). Moderate reliability was found for Bath Ankylosing Spondylitis Disease Activity Index (ICC 0.72, SDC 1.1), patient global assessment (ICC 0.58, SDC 1.5), total back pain (ICC 0.64, SDC 1.3), back pain at night (ICC 0.67, SDC 1.3), morning stiffness (ICC 0.52–0.63, SDC 1.5–2.2), fatigue (ICC 0.65, SDC 1.3) and ASAS-Health Index (ICC 0.74, SDC 2.5). Reliability and SDC for the radiographic and non-radiographic axSpA subgroups were similar.

Conclusion Overall reliability was good, and comparable levels of reliability were found for patients with radiographic and non-radiographic axSpA, even though most instruments were developed for radiographic axSpA. Composite measures showed higher reliability than single-item measures in assessing disease activity in patients with axSpA.

Key messages

What is already known about this subject?

- Most instruments used to assess effectiveness of treatment in axial spondyloarthritis were developed for and validated in patients with radiographic axial spondyloarthritis.

What does this study add?

- Overall reliability of the investigated instruments was good for all patients with axial spondyloarthritis (ie, radiographic and non-radiographic).
- Smallest detectable change of the investigated instruments was comparable between patients with radiographic and non-radiographic axial spondyloarthritis.

How might this impact on clinical practice or further developments?

- Though most instruments were developed for radiographic axial spondyloarthritis, they are also reliable for non-radiographic axial spondyloarthritis

INTRODUCTION

Uniformity in reporting primary outcomes of clinical trials allows for a direct comparison between studies investigating different therapies in the same patient population. Herein, there is an essential role for core outcome sets (COS), which contain the mandatory outcomes (domains) that should be assessed and reported as a minimum in all trials.[1 2] Over time, new instruments to assess these domains may be developed and also more data may become available regarding measurement properties of already existing instruments, underlining the need to periodically review COS. Currently, the Assessment of Spondyloarthritis international Society (ASAS) is working on an update of the original ASAS/Outcome Measures in Rheumatology (OMERACT) core set for ankylosing spondylitis (AS) of which the domains have been selected and endorsed.[3 4] An important aspect that led to this decision was that AS belongs to a broader disease spectrum, axial spondyloarthritis (axSpA), which includes two forms—that can also be regarded as two

stages- of the same disease: radiographic axSpA (r-axSpA, traditionally known as AS, that is, axSpA with definite sacroiliitis according to the modified New York (mNY) criteria[5]) and non-radiographic axSpA (nr-axSpA, that is, axSpA without definite sacroiliitis on radiographs[6]). Even though both nr-axSpA and r-axSpA are now considered part of the same disease spectrum, most instruments used to assess effectiveness of treatment were developed for and tested only in patients with r-axSpA.

The updated COS should be applicable to all patients with axSpA. Therefore, all instruments should have good psychometric properties for patients in both disease subgroups (ie, r-axSpA and nr-axSpA) to be included as mandatory instruments.[1 2] The psychometric properties include truth (domain match, face and content validity), feasibility, construct validity and discrimination (test–retest reliability, responsiveness, clinical trial discrimination and thresholds of meaning).[7] In this manuscript, we evaluate only one aspect in detail, namely test–retest reliability. Reliability is an important psychometric property, as it informs users whether the same result will be obtained if assessed twice in a situation where there is no change. Hence, the aim of this study was to assess test–retest reliability of the candidate instruments for the selected mandatory domains of the core outcome set that should be assessed in all trials evaluating a new treatment in patients with r-axSpA and nr-axSpA.[4]

METHODS

Study population

For this study, we used screening and baseline data from three large samples in axSpA: data from COAST-V and COAST-X (initiated by Eli Lilly and Company and registered with ClinicalTrials.gov as NCT02696785 and NCT02757352 respectively) and RAPID-axSpA (initiated by UCB Pharma and registered with ClinicalTrials.gov as NCT01087762). These randomised controlled trials (RCTs) are described in detail elsewhere.[8–10] In brief, all RCTs included patients aged ≥18 years who fulfilled ASAS criteria for axSpA[11] and had an inadequate response to nonsteroidal anti-inflammatory drugs (NSAIDs) or a history of intolerance to NSAIDs. COAST-V included patients with r-axSpA[8] (ie, with sacroiliitis according to the mNY criteria[5]) while COAST-X included patients with nr-axSpA[9]; and RAPID-axSpA comprised patients with either r-axSpA or nr-axSpA.[10] As these patients were entering an RCT, they needed to have active disease at screening and baseline, defined as a Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)[12] score of ≥4 and total back pain in the past week ≥4 (on a 0–10 Numeric Rating Scale (NRS)).

Outcomes

The ASAS-OMERACT core domain set for axSpA[4] describes the domains that should be measured in axSpA trials investigating symptom modifying and disease-modifying therapies. Seven domains are mandatory in all axSpA trials: disease activity, pain, morning stiffness, fatigue, physical function, overall functioning and health and adverse events. Information from all the instruments (n=13) employed to assess these domains -with the exception of adverse events- at both screening and baseline in COAST-V, COAST-X and RAPID-axSpA was used to evaluate test–retest reliability of each instrument.

Four instruments that could be used to assess the domain disease activity were available: the Ankylosing Spondylitis Disease Activity Score (ASDAS) -specifically ASDAS-C reactive protein (CRP),[13] the BASDAI using NRS answer modalities,[12] the patient global assessment (PtGA) using an NRS[14] and CRP, measured in mg/L. Two of the instruments used to assess pain were available: 0–10 NRS for total back pain in the past week and 0–10 NRS for pain at night in the past week.[14] Questions 5 (How would you describe the overall level of morning stiffness you have had from the time you wake up?) and 6 (How long does your morning stiffness last from the time you wake up?) of the BASDAI and a composite score of questions 5 and 6 ((Q5 +Q6)/2) were the instruments available to evaluate morning stiffness. The one instrument available to estimate fatigue was question 1 of the BASDAI. To evaluate physical function, one instrument was present: the Bath Ankylosing Spondylitis Functional Index (BASFI).[15] Two of the instruments that could survey overall functioning and health were available: the ASAS-Health Index (ASAS-HI)[16] and Medical Outcomes Study 36-item Short-Form Health Survey (SF-36).[17] All these instruments are commonly used in trials assessing treatment effect in axSpA and have shown content, face and construct validity.[18]

Spinal mobility was considered an important but optional domain in the axSpA ASAS/OMERACT domain core set.[4] Nonetheless, it was included in this study as it is often assessed in clinical trials and daily practice. One composite instrument and two additional single measures that can be used to evaluate spinal mobility were evaluated: the Bath Ankylosing Spondylitis Metrology Index (BASMI) linear[19] (including modified Schober, lateral spinal flexion, tragus-to-wall distance, cervical rotation, intermalleolar distance) and chest expansion and occiput-to-wall distance.[14]

Statistical analyses

Test–retest reliability was assessed by intraclass correlation coefficient (ICC) (two-way random effect model with absolute agreement[20 21]). An ICC >0.9 was an indication of excellent reliability, >0.75 to 0.9 of good reliability, 0.5 to 0.75 of moderate reliability and ICC <0.5 of poor reliability.[21] Bland and Altman plots were created for each instrument to assess mean difference and 95% limits of agreement and to evaluate homoscedasticity. Measurement error as a measure of the scale was assessed by analysing the smallest detectable change (SDC) based on the 95% limits of agreement using the formula: SDC=1.96×SD of the mean difference of the two assessments/($\sqrt{2}$x $\sqrt{2}$).[22] The SDC corresponds to the

minimum change beyond measurement error that can be detected in an individual patient over time with 95% likelihood. Calculation of the limits of agreement (and the SDC) assumed that reliability was homoscedastic.

In this study, we operated under an a priori assumption underlying the test–retest experiments, namely that in truth the scores for all instruments do not change over the limited period of time between assessments (ie, there is no systematic error). This assumption of no change has been proven by the Bland and Altman plots, which demonstrated that the mean difference between test and retest was always (very close to) zero, indicating that the no systematic error assumption holds.

As there was a large variation in the number of days between screening and baseline assessments in both datasets, it was decided to use a maximum time interval of 28 days between both visits as a cut-off for inclusion in this study.

Unfortunately, in the RAPID-axSpA dataset the PtGA was only assessed at baseline, and the baseline values were used to calculate ASDAS both at screening and baseline. As the ASDAS is calculated from the PtGA, questions 2, 3 and 6 from the BASDAI and CRP,[13] the results of this dataset should be interpreted with caution, as variability in patient global was not considered and as a result the reliability of the ASDAS may be artificially improved. However, the values in the COAST trials were very similar.

Results were bundled per domain and presented for all axSpA patients, followed by information per disease subgroup (ie, r-axSpA and nr-axSpA). Data from both COAST datasets were combined to assess test–retest reliability of the instruments in axSpA patients.

RESULTS

A total of 341 r-axSpA patients in the COAST-V dataset, 302 nr-axSpA patients in the COAST-X dataset and 326 patients (177 r-axSpA and 149 nr-axSpA) in the RAPID-axSpA dataset had data available at screening and baseline. From these, 104 r-axSpA patients from COAST-V, 104 nr-axSpA patients from COAST-X and 221 patients from RAPID-axSpA (119 r-axSpA and 102 nr-axSpA) who had both measurements for at least one of the assessed instruments within a time frame of 28 days were included in this analysis.

Of the included r-axSpA patients from COAST-V 81% were male median (IQR) age was 39 (34–47) and mean (SD) symptom duration 15.1 (9.9) years. The selection of nr-axSpA patients from COAST-X included 55% male patients, with a median age of 38 (27–49) and mean symptom duration of 9.9 (8.8) years. In RAPID-axSpA 62% of the included patients were male (74% in r-axSpA, 49% in nr-axSpA), the median age range was 31–35 years (46–50 in r-axSpA, 31–35 in nr-axSpA) and mean symptom duration was 6.0 (6.9) years (7.4 (7.6) in r-axSpA, 4.3 (5.6) in nr-axSpA).

The mean symptom duration in the patient selection included in this study was somewhat shorter than the mean symptom duration of the entire study populations (COAST-V 16.1 (10.9); COAST-X 10.7 (9.7); RAPID-axSpA 6.7 (7.4)). Median age and the percentage of female patients were similar to the original study populations.[8–10]

The number of days between assessments ranged between 8 and 28 days in COAST-V, between 9 and 28 days in COAST-X and between 2 and 28 days in RAPID-axSpA; the mean (SD) number of days between assessments were 22 (5) in COAST-V, 21 (5) in COAST-X and 18 (7) days in RAPID-axSpA. The proportion of missing data varied somewhat between measurements and datasets, but was always very small (<5%). Participants with missing data for an instrument at either screening or baseline were excluded from analysis for that specific instrument. The number of available data per instrument is provided in table 1. Information available from the literature regarding reliability of the instruments included in the current study is presented in table 1.[23–36]

Detailed results from all trials and subgroups are provided in tables 1 and 2. In the text, reliability per domain is described only for the total axSpA group in the COAST datasets, as these included most instruments. Only if reliability varied considerably between subgroups or trials, reliability of these groups is discussed additionally.

Regarding the four instruments assessing disease activity: good reliability was found for ASDAS (ICC 0.79, SDC 0.6) and CRP in COAST (ICC 0.79, SDC 12.3), whereas reliability for CRP in the RAPID-axSpA dataset was slightly lower (ICC 0.72, SDC 17.0) (table 1). Reliability was moderate for BASDAI (ICC 0.72, SDC 1.1); and for the PtGA reliability was moderate (ICC 0.58, SDC 1.5) too, except for the r-axSpA group, for which reliability was poor (ICC 0.48, SDC 1.6). The two instruments used to evaluate pain showed moderate reliability (NRS total back pain (ICC 0.64, SDC 1.3); NRS back pain at night (ICC 0.67, SDC 1.3)). Moderate reliability was found for the instruments used to assess morning stiffness (ICC 0.52–0.63, SDC 1.5–2.2) as well. The instrument used to determine fatigue showed moderate reliability (ICC 0.65, SDC 1.3). The data showed good reliability (ICC 0.87, SDC 1.1) for the BASFI, used to measure physical function. For the two instruments used to survey overall functioning and health, good reliability was found for the Physical Component Summary (ICC 0.81, SDC 4.7) and Mental Component Summary (ICC 0.80, SDC 7.3) subscales of the SF-36, and the ASAS-HI had moderate reliability (ICC 0.74, SDC 2.5), except for the nr-axSpA subgroup in which reliability was good (ICC 0.77, SDC 2.5). In the domain spinal mobility, reliability was excellent (ICC 0.93, SDC 0.6) for BASMI in RAPID-axSpA. Tragus-to-wall and occiput-to-wall distance showed excellent reliability, except for the nr-axSpA subpopulation,

Table 1 Test–retest data of assessed instruments in COAST (combined data COAST-V & COAST-X) and RAPID-axSpA, 28-day interval

Data source		N	Screening mean (SD)	Baseline mean (SD)	Mean difference (95% CI)	ICC (95% CI)	SDC	Data from literature	
								ICC	MCID/MCII/SDC
Disease activity								Disease activity	
	ASDAS (0.6 to 8)							ASDAS	
COAST	axSpA	204	3.8 (.9)	3.8 (.8)	.08 (.00 to .16)	.79 (.73 to .84)	0.6	ICC: 0.95[32]	SDC range 1.01–1.18; MCII: 1.1[29]
	r-axSpA	103	3.7 (.9)	3.7 (.8)	.03 (–0.08 to .13)	.80 (.71 to .86)	0.6		
	nr-axSpA	101	4.0 (.9)	3.8 (.9)	.13 (.02 to .24)	.78 (.69 to .85)	0.6		
	ASDAS (0.6 to 8)*								
RAPID-axSpA	axSpA	215	4.0 (.8)	4.0 (.9)	.01 (–0.06 to .08)	.79 (.73 to .83)	0.5		
	r-axSpA	118	4.1 (.7)	4.0 (.9)	.01 (–0.09 to .12)	.75 (.66 to .82)	0.6		
	nr-axSpA	97	3.9 (.8)	3.9 (.9)	.01 (–0.09 to .11)	.83 (.76 to .88)	0.5		
	BASDAI (0 to 10)							BASDAI	
COAST	axSpA	208	6.9 (1.5)	6.9 (1.4)	.05 (–0.10 to .20)	.72 (.65 to .78)	1.1	ICC range 0.87–0.94[23 30 34]	MCID: 1.3[34]; MCII: 1.1–1.2[27] SDC: 0.9[30]
	r-axSpA	104	6.6 (1.4)	6.6 (1.3)	.05 (-. 16 to .26)	.67 (.55 to .77)	1.1		
	nr-axSpA	104	7.2 (1.5)	7.2 (1.5)	.05 (–0.16 to .26)	.74 (.64 to .82)	1.1		
	BASDAI (0 to 10)								
RAPID-axSpA	axSpA	217	6.5 (1.5)	6.6 (1.5)	–0.13 (–0.31 to .05)	.62 (.53 to .70)	1.3		
	r-axSpA	119	6.5 (1.5)	6.6 (1.6)	–0.11 (–0.36 to .15)	.61 (.49 to .71)	1.3		
	nr-axSpA	98	6.6 (1.4)	6.7 (1.5)	–0.17 (–0.42 to .09)	.64 (.50 to .74)	1.2		
	Patient global (0 to 10)							Patient global	
COAST	axSpA	208	7.0 (1.7)	7.1 (1.6)	–0.05 (–0.25 to .16)	.58 (.48 to .66)	1.5	ICC range 0.91–0.93[23 31]	MCII: 1.4[33] SDC†: 1.8[23]
	r-axSpA	104	6.7 (1.6)	6.8 (1.5)	–0.13 (–0.44 to .19)	.48 (.32 to .61)	1.6		
	nr-axSpA	104	7.3 (1.6)	7.3 (1.7)	.03 (–0.24 to .30)	.64 (.51 to .74)	1.3		
	CRP (mg/dL)								
COAST	axSpA	204	16.7 (21.7)	13.8 (17.8)	2.93 (1.20 to 4.67)	.79 (.73 to .84)	12.3		
	r-axSpA	103	15.7 (19.0)	14.3 (17.2)	1.39 (–0.46 to 3.24)	.86 (.80 to .90)	9.3		
	nr-axSpA	101	17.8 (24.2)	13.3 (18.6)	4.51 (1.56 to 7.46)	.75 (.63 to .83)	14.6		
	CRP (mg/dL)								
RAPID-axSpA	axSpA	219	20.6 (20.9)	20.4 (25.6)	.18 (-2.17 to 2.53)	.72 (.65 to .78)	17.0		
	r-axSpA	118	21.1 (18.8)	22.2 (28.1)	–1.17 (-5.12 to 2.77)	.60 (.47 to .70)	21.0		
	nr-axSpA	101	20.1 (23.1)	18.4 (22.2)	1.76 (–0.37 to 3.90)	.89 (.84 to .92)	10.5		
Pain								Pain	
	Total back pain (0 to 10)							Total back pain	
COAST	axSpA	208	7.1 (1.6)	7.2 (1.5)	–0.15 (–0.32 to .03)	.64 (.56 to .72)	1.3	ICC range 0.86–0.9[23 34]	MCID: 1.6 (range 1.5–1.6)[34] SDC: 1.8[30]
	r-axSpA	104	6.9 (1.5)	7.0 (1.3)	–0.13 (–0.38 to .13)	.58 (.43 to .69)	1.3		
	nr-axSpA	104	7.2 (1.6)	7.4 (1.6)	–0.17 (–0.42 to .07)	.69 (.57 to .78)	1.3		

Continued

Table 1 Continued

Data source		N	Screening mean (SD)	Baseline mean (SD)	Mean difference (95% CI)	ICC (95% CI)	SDC	Data from literature: ICC	Data from literature: MCID/MCII/SDC
	Night pain (0 to 10)							Night pain	
COAST	axSpA	208	7.0 (1.8)	7.1 (1.6)	−0.07 (−0.26 to .12)	.67 (.59 to .74)	1.3	ICC range 0.83–0.92[23 34]	MCID: 1.8 (range 1.5–2.1)[34]
	r-axSpA	104	6.9 (1.8)	6.8 (1.6)	.05 (−0.22 to .32)	.65 (.53 to .75)	1.3		
	nr-axSpA	104	7.2 (1.8)	7.3 (1.7)	−0.19 (−0.47 to .08)	.69 (.56 to .78)	1.4		
Morning stiffness								Morning stiffness	
	BASDAI Q5: Morning stiffness severity (0 to 10)							BASDAI Q5: Morning stiffness severity	
COAST	axSpA	208	7.3 (1.9)	7.2 (1.7)	.13 (-3.02 to 3.29)	.63 (54 to .70)	1.5	ICC 0.85[36]	SDC†: 1.4[36]
	r-axSpA	104	7.2 (1.8)	6.9 (1.6)	.26 (-2.61 to 3.13)	.64 (.51 to .74)	1.4		
	nr-axSpA	104	7.5 (2.1)	7.5 (1.8)	.01 (-3.40 to 3.42)	.62 (.48 to .72)	1.7		
	BASDAI Q6: Morning stiffness duration (0 to 10)‡								
COAST	axSpA	208	6.2 (2.4)	6.2 (2.2)	−0.01 (-4.55 to 4.52)	.52 (.41 to .61)	2.2		
	r-axSpA	104	6.0 (2.4)	5.8 (2.2)	.13 (-4.43 to 4.68)	.51 (.35 to .64)	2.2		
	nr-axSpA	104	6.4 (2.4)	6.5 (2.3)	−0.15 (-4.68 to 4.37)	.52 (.37 to .65)	2.2		
	BASDAI Morning stiffness (0 to 10) composite(Q5 +Q6/2)							BASDAI Morning stiffness composite	
COAST	axSpA	208	6.7 (1.8)	6.7 (1.7)	.06 (−0.15 to .27)	.63 (.55 to .71)	1.5	ICC range 0.85–0.91[34 36]	MCID: 1.7 (range 1.0–2.7)[34]
	r-axSpA	104	6.6 (1.8)	6.4 (1.6)	.19 (−0.11 to .49)	.58 (.43 to .69)	1.5		
	nr-axSpA	104	6.9 (1.9)	7.0 (1.8)	−0.07 (−0.37 to .22)	.67 (.55 to .77)	1.5		
	BASDAI Morning stiffness (0 to 10) composite(Q5 +Q6/2)								
RAPID-axSpA	axSpA	217	6.3 (2.6)	6.3 (2.4)	.01 (−0.28 to .32)	.60 (.51 to .68)	2.2		
	r-axSpA	119	6.5 (2.5)	6.4 (2.4)	.09 (−0.31 to .50)	.60 (.47 to .70)	2.2		
	nr-axSpA	89	6.2 (2.7)	6.3 (2.4)	−0.08 (−0.53 to .37)	.61 (.47 to .72)	2.2		
Fatigue								Fatigue	
	BASDAI Q1: Fatigue (0 to 10)							BASDAI Q1	
COAST	axSpA	208	7.2 (1.7)	7.1 (1.6)	.08 (−0.10 to .27)	.65 (.57 to .72)	1.3	ICC: 0.60–0.85[34 35]	MCID: 1.1 (range 1.0–1.5)[34] SDC: 1.7[30]
	r-axSpA	104	6.8 (1.6)	6.8 (1.6)	.01 (−0.27 to .29)	.59 (.45 to .71)	1.4		
	nr-axSpA	104	7.5 (1.7)	7.4 (1.6)	.16 (−0.09 to .41)	.68 (.57 to .77)	1.3		
	BASDAI Q1: Fatigue (0 to 10)								
RAPID-axSpA	axSpA	217	6.5 (1.9)	6.8 (1.9)	−0.29 (−0.54 to −0.03)	.53 (.42 to .62)	1.8		
	r-axSpA	119	6.6 (1.9)	6.8 (2.0)	−0.13 (−0.48 to .21)	.54 (.40 to .65)	1.8		
	nr-axSpA	98	6.4 (1.9)	6.9 (1.8)	−0.47 (−0.84 to −0.10)	.51 (.35 to .64)	1.8		
Physical function								Physical function	
	BASFI (0 to 10)							BASFI	

Continued

Table 1 Continued

Data source		N	Screening mean (SD)	Baseline mean (SD)	Mean difference (95% CI)	ICC (95% CI)	SDC	Data from literature	
								ICC	MCID/MCII/SDC
COAST	axSpA	208	6.2 (2.1)	6.3 (2.0)	−0.13 (−0.27 to .02)	.87 (.83 to .90)	1.1	ICC range 0.92–0.94[15 30]	MCID: 1.1 (range 1.0–1.1)[34] MCII: 0.6–1.1[27 33] SDC: 0.7[30]
	r-axSpA	104	6.0 (2.1)	6.0 (2.1)	−0.08 (−0.27 to .11)	.89 (.84 to .92)	0.9		
	nr-axSpA	104	6.4 (2.0)	6.5 (2.0)	−0.17 (−0.39 to .05)	.84 (.77 to .89)	1.1		
Overall functioning and health								Overall functioning & health	
	ASAS Health Index (0–17)							ASAS Health Index	
COAST	axSpA	208	8.6 (3.6)	8.7 (3.6)	−0.07 (−0.43 to .28)	.74 (.68 to .80)	2.5	ICC range 0.84–0.98[24–26 28]	SDC: 3.0[25]
	r-axSpA	104	7.7 (3.4)	7.7 (3.3)	−0.05 (−0.57 to .48)	.68 (.56 to .77)	2.6		
	nr-axSpA	104	9.5 (3.7)	9.8 (3.7)	−0.10 (−0.59 to .39)	.77 (.68 to .84)	2.5		
	SF-36 PCS (0–100)							SF-36 PCS	
COAST	axSpA	208	34.0 (7.9)	34.9 (7.8)	−0.89 (-1.55 to −0.23)	.81 (.75 to .85)	4.7		MCID: 3.8[34]
	r-axSpA	104	35.7 (8.0)	36.3 (7.4)	−0.63 (-1.49 to .23)	.83 (.76 to 88)	4.3		
	nr-axSpA	104	32.4 (7.6)	33.5 (8.1)	−1.15 (-2.17 to −0.14)	.77 (.68 to .84)	5.1		
	SF-36 MCS (0–100)							SF-36 MCS	
COAST	axSpA	208	47.8 (11.7)	48.0 (11.8)	−0.15 (-1.17 to .87)	.80 (.75 to .84)	7.3		MCID: 2.4[34]
	r-axSpA	104	50.5 (10.3)	50.4 (10.5)	.08 (-1.26 to 1.42)	.78 (.69 to .85)	6.7		
	nr-axSpA	104	45.2 (12.6)	45.6 (12.5)	−0.37 (-1.93 to 1.18)	.80 (.72 to .86)	7.8		

ASAS, Assessment of Spondyloarthritis international Society Health Index scores range from 0 to 17, higher scores indicate worse health status; ASDAS, Ankylosing Spondylitis Disease Activity Score (scores range from 0.6 to 8, determined by the level of CRP or ESR (8 is the approximate maximum, given a CRP of 200), higher scores signify higher disease activity); axSpA, axial spondyloarthritis; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index (scores range from 0 to 10, higher scores signify higher disease activity); BASFI, Bath Ankylosing Spondylitis Functional Index (scores range from 0 to 10, higher scores signify greater impairment); COAST-V, a randomised placebo-controlled trial assessing efficacy and safety of ixekizumab for patients with radiographic axSpA; COAST-X, a randomised placebo-controlled trial assessing efficacy and safety of ixekizumab for patients with non-radiographic axSpA; CRP, C reactive protein (measured in mg/dL); ICC, intraclass correlation coefficient; MCID, minimal clinically important difference; MCII, minimal clinically important improvement; Night pain, Nocturnal back pain in the past week measured using an NRS scale (scores range from 0 to 10, higher scores indicating more pain); nr-axSpA, non-radiographic axSpA; Patient global, PtGA of disease activity measured using an NRS scale (scores range from 0 to 10, higher scores indicate worse health); RAPID-axSpA, a randomised placebo-controlled trial assessing efficacy and safety of certolizumab pegol for patients with radiographic and non-radiographic axSpA; r-axSpA, radiographic axSpA; SDC, smallest detectable change; SDD, SD of the mean difference of the two assessments; SF-36 PCS, Physical Component Summary of the Medical Outcomes Study 36-item Short-Form Health Survey (scores range from 0 to 100, higher scores indicate better health); SF-36 MCS, Mental Component Summary of the Medical Outcomes Study 36-item Short-Form Health Survey (scores range from 0 to 100, higher scores indicate better health); Total back pain, in the past week measured using an NRS scale (scores range from 0 to 10, higher scores indicating more pain).

*PtGA was only assessed at baseline, and baseline values were used to calculate ASDAS at both time points, meaning variability in PtGA was not considered and reliability of the ASDAS may be artificially improved.

†Calculated from the SDD using the formula $SDC=1.96\times SDD/(\sqrt{2}\times\sqrt{2})$.

‡Score range 0–10, with three anchors: '0 hours' (score 0), '1 hour' (score 5) and '2 or more hours' (score 10).

ASAS, Assessment of Spondyloarthritis international Society; ASDAS, Ankylosing Spondylitis Disease Activity Score; axSpA, axial spondyloarthritis; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; CRP, C reactive protein; ICC, intraclass correlation coefficient; MCID, minimal clinically important difference; MCII, minimal clinically important improvement; MCS, Mental Component Summary; nr-axSpA, non-radiographic axSpA; NRS, Numeric Rating Scale; PCS, Physical Component Summary; PtGA, patient global assessment; r-axSpA, radiographic axSpA; SDC, smallest detectable change; SDD, SD of the mean difference; SF-36, 36-item Short-Form Health Survey.

Table 2 Test–retest data of spinal mobility instruments measured in RAPID-axSpA, 28-day interval

	N	Screening mean (SD)	Baseline mean (SD)	Mean difference (95% CI)	ICC (95% CI)	SDC	*Data from literature* ICC	MCID/MCII/SDC
BASMI (0 to 10)							BASMI	
axSpA	218	3.6 (1.5)	3.7 (1.6)	−0.12 (−0.20 to −0.05)	.93 (.91 to .95)	0.6	ICC rage 0.91–0.97[37 39 41]	SDC range 0.82–0.95[37 41]
r-axSpA	117	4.2 (1.5)	4.4 (1.5)	−0.19 (−0.31 to −0.08)	.91 (.87 to .94)	0.6		
nr-axSpA	101	3.0 (1.3)	3.0 (1.4)	−0.05 (−0.15 to .06)	.93 (.90 to .95)	0.5		
Modified Schober (cm)							Modified Schober	
axSpA	216	3.9 (2.2)	3.8 (2.0)	.12 (−0.03 to .27)	.86 (.82 to .89)	1.1	ICC inter-observer range 0.75–0.96[38 40 42 43]; ICC intra-observer range 0.63–0.94[38 40 43]	SDC 1.4[40]
r-axSpA	116	3.5 (2.1)	3.3 (2.0)	.15 (−0.05 to .36)	.85 (.79 to .90)	1.1		
nr-axSpA	100	4.4 (2.3)	4.3 (1.8)	.09 (−0.14 to .31)	.85 (.78 to .90)	1.1		
Lateral spinal flexion (cm)							Lateral spinal flexion	
axSpA	216	12.6 (6.6)	12.1 (6.0)	.39 (−0.02 to .79)	.80 (.75 to .85)	2.9	ICC inter-observer range 0.–0.77-0.98[38 40 42 43]; ICC intra-observer range 0.65–0.98[38 40 43]	SDC 5.1[40]
r-axSpA	116	11.6 (7.3)	10.3 (5.6)	.98 (.46 to 1.50)	.75 (.65 to .82)	2.8		
nr-axSpA	100	13.8 (5.6)	14.1 (5.8)	−0.30 (−0.91 to .32)	.86 (.79 to .90)	3.0		
Tragus-to-wall distance (cm)							Tragus-to-wall distance	
axSpA	218	13.5 (4.9)	13.7 (5.0)	−0.19 (−0.46 to .07)	.92 (.90 to .94)	1.9	ICC inter-observer range 0.85–0.98[38 42 43]; ICC intra-observer range 0.94–0.98[38 43]	
r-axSpA	117	14.8 (5.8)	14.9 (5.7)	−0.14 (−0.45 to .17)	.96 (.94 to .97)	1.6		
nr-axSpA	101	12.1 (3.0)	12.4 (3.7)	−0.25 (CI −0.70 to .19)	.78 (.70 to .85)	2.2		
Cervical rotation (degrees)							Cervical rotation	
axSpA	218	55.9 (20.9)	54.7 (21.4)	1.20 (−0.38 to 2.79)	.85 (.80 to .88)	11.5	ICC inter-observer range 0.69–0.94[38 40 42]; ICC intra-observer range 0.56–0.95[38 40]	SDC 12.2[40]
r-axSpA	117	48.5 (19.7)	46.6 (19.7)	1.89 (−0.62 to 4.41)	.76 (.67 to .83)	13.4		
nr-axSpA	101	64.4 (19.2)	64.0 (19.3)	.40 (-1.40 to 2.21)	.89 (.84 to .92)	8.9		
Intermalleolar distance (cm)							Intermalleolar distance	
axSpA	216	98.1 (25.8)	96.6 (27.3)	1.53 (−0.10 to 3.16)	.89 (.87 to .92)	11.7	ICC inter-observer 0.93[42]; ICC intra-observer 0.72[40]	SDC 20.2[40]
r-axSpA	116	94.5 (25.9)	93.2 (28.2)	1.40 (-1.06 to 3.85)	.88 (.83 to .92)	12.9		
nr-axSpA	100	102.2 (25.2)	100.6 (25.7)	1.69 (−0.39 to 3.76)	.92 (.88 to .94)	10.2		
Chest expansion (cm)							Chest expansion	
axSpA	217	3.7 (2.2)	3.7 (2.0)	−0.06 (−0.22 to .10)	.78 (.72 to .83)	1.1	ICC inter-observer range 0.55–0.85([40 42 43]; ICC intra-observer range 0.63–0.95[40 43]	SDC 2.2[40]
r-axSpA	118	3.6 (2.5)	3.4 (2.0)	.07 (−0.14 to .29)	.76 (.68 to .83)	1.1		
nr-axSpA	99	3.9 (1.8)	4.2 (2.0)	−0.22 (−0.46 to .01)	.81 (.72 to .87)	1.1		
Occiput-to-wall distance (cm)							Occiput-to-wall distance	
axSpA	218	3.7 (5.6)	3.6 (5.5)	.06 (−0.19 to .30)	.95 (.93 to .96)	1.8	ICC inter-observer range 0.84–0.89([40 42 43]; ICC intra-observer range 0.49–0.94[40 43]	SDC 0.9[40]
r-axSpA	118	4.5 (6.3)	4.7 (6.4)	−0.11 (−0.41 to .19)	.97 (.96 to .98)	1.6		
nr-axSpA	100	2.6 (4.5)	2.4 (3.7)	.25 (−0.14 to .65)	.88 (.83 to .92)	1.9		

BASMI, Bath Ankylosing Spondylitis Metrology Index (scores range from 0 to 10, higher scores signify greater impairment); ICC, intraclass correlation coefficient; MCID, minimal clinically important difference; MCII, minimal clinically important improvement; nr-axSpA, non-radiographic axSpA; RAPID-axSpA, a randomised placebo-controlled trial assessing efficacy and safety of certolizumab pegol for patients with radiographic and non-radiographic axSpA; r-axSpA, radiographic axSpA; SDC, smallest detectable change; SDD, SD of the mean difference of the two assessments

axSpA, axial spondyloarthritis; BASMI, Bath Ankylosing Spondylitis Metrology Index; ICC, intraclass correlation coefficient; MCID, minimal clinically important difference; MCII, minimal clinically important improvement; nr-axSpA, non-radiographic axSpA; r-axSpA, radiographic axSpA; SDC, smallest detectable change; SDC, smallest detectable change; SDD, SD of the mean difference.

for which the reliability was good. For all other mobility measures reliability was good (table 2).[37–43]

Bland and Altman plots showed a reasonably homoscedastic variation for all measurement instruments, with the exception of CRP where the variation was more pronounced in the lower end of the range (online supplemental figures 1–27).

DISCUSSION

The results from this study showed that the test–retest reliability of the investigated instruments was moderate to excellent and similar in the axSpA group and each of the disease subgroups r-axSpA and nr-axSpA. Furthermore, for those instruments where data was available from the COAST and RAPID-axSpA studies, levels of reliability were comparable between datasets as well. Finally, we found ICCs were higher for multi-item instruments compared with single-item instruments in the same domain. This is reasonable as the impact of variance caused by measurement error in the individual items of a multi-item instrument is reduced when they are combined into a single score, resulting in a more precise score for a multi-item instrument compared with its single-item counterparts.[44 45]

For all instruments assessed in this study, ICCs were somewhat lower than those previously reported in the literature, with the exception of the spinal mobility measures. This is not unexpected as all patients included in this study had high disease activity, which resulted in less variability in scores between patients for the investigated instruments (eg, BASDAI and total back pain had a possible range of 4–10 instead of 0–10). It has been shown that reduced variability in scores decreases ICCs in case of unchanged number of observations and measurement error.[21 46] This might explain why for almost all measurement instruments the reliability found in this study was somewhat lower than those reported previously. Other characteristics, such as the proportion of female patients, age and symptom duration of the patients included in this study were comparable to the populations included in previous studies investigating reliability.[23 25 27 29 30 32–35] The decreased variability in scores has an opposite effect on the SDCs, as the mean difference between two assessments (and its SD) is expected to be smaller when the scoring range is reduced, this applies to scores between patients as well as between two measurements within the same patient. An SDC represents a minimum that can be observed reliably based on measurement error. This can be compared with a minimal clinically important improvement (MCII, defined in relation to an external standard for an individual patient) and minimal clinically important difference (MCID, defined by an external standard between (groups of) patients). We compared the observed SDCs with the published SDCs, MCIIs and MCIDs in the literature. The SDCs for ASDAS found in this study were indeed lower than the MCII defined in the literature,[29] while SDCs for BASDAI, PtGA and BASFI found in these datasets were similar to the previously reported MCIIs.[27 33] Based on the data analysed in this study, we can conclude ASDAS has the best reliability and smallest SDC of the instruments used to assess disease activity.

For total back pain and pain at night in the past week, SDCs were smaller than the MCID defined in the literature,[34] and ICCs were comparable for both instruments. The data for the fatigue and stiffness questions of the BASDAI was inconclusive. In the COAST-X and COAST-V datasets SDCs were similar to the reported MCIDs.[34 47–49] Conversely, measurement error in the RAPID-axSpA was somewhat larger, complicating detection of the MCID. Comparing the ICCs and SDCs of the various instruments used to assess morning stiffness in the COAST datasets, duration of morning stiffness seems slightly less reliable compared with severity of morning stiffness and the composite score. Finally, the SDC for the ASAS-HI was slightly smaller than previously reported,[25] which could be the result of the afore mentioned limited range in disease activity in the current study populations. Compared with the SF-36, the SDC of the ASAS-HI was higher (12% vs 5%–7% of the total score range) and the ICC slightly lower, indicating the SF-36 might have better reliability. However, the ASAS-HI is a disease-specific instrument, whereas the SF-36 is a general instrument, thus other measurement properties are vital for a final conclusion. Before a definite decision can be made regarding which instrument is best to assess each domain, the other measurement properties will have to be collected too.

This study used data from three recent trials in axSpA, which ensured all instruments currently used in clinical trials were represented. All patients included in these datasets had active disease and were candidate to receive a disease-modifying therapy, which matches the target group of the ASAS-OMERACT core outcome set.[4] As the core outcome set will be used in clinical trials assessing the effect of treatment in axSpA and RCTs in principle require patients with active disease, the data from this study provide valuable information on the reliability of measurement instruments in this patient group. Furthermore, an equal number of patients with r-axSpA and nr-axSpA were included, thereby representing all patients with axSpA disease. Nonetheless, there were limitations to this study, the most important one being the relatively long time-interval used in the current study to ensure the sample sizes would be large enough, which might explain some of the differences found between the literature and the results in this study. Based on the data from this study and information available in the literature, ASDAS, BASDAI, PtGA and CRP are reliable measures to assess disease activity in all patients with axSpA, both total back pain and pain at night in the past week could be considered reliable in assessing pain, questions 5 and 6 of the BASDAI can be used to reliably assess morning stiffness, BASDAI question 1 can reliably evaluate fatigue, BASFI was found reliable to investigate physical functioning,

ASAS-HI and SF-36 were found reliable to survey overall functioning & health, and BASMI and its components as well as chest expansion can be used to reliably assess spinal mobility. Further research will have to focus on collecting information on the other psychometric properties before a definite decision can be made regarding the best instrument for each domain.

CONCLUSION

The results from this study showed overall reliability was good and levels of reliability were comparable for patients with r-axSpA and nr-axSpA, indicating ASDAS, BASDAI, PtGA, CRP, NRS total back pain, NRS back pain at night, BASFI, ASAS-HI, SF-36 and BASMI are reliable measures for all patients with axSpA, even though most instruments were developed for r-axSpA. Composite measures showed higher reliability than single-item measures in assessing disease activity and spinal mobility in patients with axSpA and may therefore be preferred over single-item instruments for this aspect of the OMERACT filter.

Acknowledgements This publication is based on research using data from UCB Pharma that has been made available through Vivli. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication. Eli Lilly and Company (Indianapolis, IN, USA) provided the data from COAST-V and COAST-X used in this study and supported this study.

Contributors All authors were involved in the planning, conduct and reporting of the work presented in this manuscript. AB accepts full responsibility for the finished work had access to the data, and controlled the decision to publish. Data from this study was kindly provided by Eli Lilly and Company and UCB Pharma, without whom this study would not have been possible.

Funding The Assessment of Spondyloarthritis international Society (ASAS) funded Anne Boel and Victoria Navarro-Compán for the project to update the core outcome set. COAST-V and COAST-X were funded by Eli Lilly and Company and the RAPID-axSpA study was funded by UCB Pharma.

Competing interests DvdH has received consulting fees from AbbVie, Amgen, Astellas, AstraZeneca, BMS, Boehringer Ingelheim, Celgene, Daiichi, Eli Lilly and Company, Galapagos, Gilead, GlaxoSmithKline, Janssen, Merck, Novartis, Pfizer, Regeneron, Roche, Sanofi, Takeda, and UCB Pharma and is director of Imaging Rheumatology BV. JC-CW has served as a consultant for Eli Lilly and Company, Pfizer, Celgene, Chugai, UCB Pharma, and TSH Taiwan; has received research grants from Bristol-Myers Squibb, Eli Lilly and Company, Janssen, Pfizer, Sanofi-Aventis, and Novartis; and has served on a speakers bureau for Abbott, Bristol-Myers Squibb, Chugai, Eisai, Janssen, and Pfizer. VN-C has received honoraria/research support from: Abbvie, BMS, Janssen, Eli Lilly, MSD, Novartis, Pfizer, Roche and UCB. AB has no competing interest to report.

Patient consent for publication Not applicable.

Ethics approval Independent Ethics Committees or Institutional Review Boards at participating sites approved the COAST-V, COAST-X and RAPID-axSpA studies, for more details we kindly refer to the original publications. All 3 trials were performed in accordance to the Good Clinical Practice guidelines and the Declaration of Helsinki and included patients provided written informed consent prior to inclusion in the respective trials.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Data for this study were kindly provided by Eli Lilly and Company and UCB Pharma, we refer any interested parties to these companies.

ORCID iDs
Anne Boel http://orcid.org/0000-0003-2016-1744
Désirée van der Heijde http://orcid.org/0000-0002-5781-158X

REFERENCES

1 Boers M, Kirwan JR, Tugwell P. *OMERACT Handbook*, 2018.
2 Williamson PR, Altman DG, Bagley H, *et al*. The comet Handbook: version 1.0. *Trials* 2017;18:280.
3 Boel A, Navarro-Compán V, Boonen A, *et al*. Domains to be considered for the core outcome set of axial spondyloarthritis: results from a 3-round Delphi survey. *J Rheumatol* 2021:jrheum.210206.
4 Navarro-Compán V, Boel A, Boonen A, *et al*. The ASAS-OMERACT core domain set for axial spondyloarthritis. *Semin Arthritis Rheum* 2021;390. doi:10.1016/j.semarthrit.2021.07.021. [Epub ahead of print: 01 Aug 2021].
5 van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the new York criteria. *Arthritis Rheum* 1984;27:361–8.
6 Rudwaleit M, Khan MA, Sieper J. The challenge of diagnosis and classification in early ankylosing spondylitis: do we need new criteria? *Arthritis Rheum* 2005;52:1000–8.
7 Beaton DE, Maxwell LJ, Shea BJ, *et al*. Instrument selection using the OMERACT filter 2.1: the OMERACT methodology. *J Rheumatol* 2019;46:1028–35.
8 van der Heijde D, Cheng-Chung Wei J, Dougados M, *et al*. Ixekizumab, an interleukin-17A antagonist in the treatment of ankylosing spondylitis or radiographic axial spondyloarthritis in patients previously untreated with biological disease-modifying anti-rheumatic drugs (COAST-V): 16 week results of a phase 3 randomised, double-blind, active-controlled and placebo-controlled trial. *Lancet* 2018;392:2441–51.
9 Deodhar A, van der Heijde D, Gensler LS, *et al*. Ixekizumab for patients with non-radiographic axial spondyloarthritis (COAST-X): a randomised, placebo-controlled trial. *Lancet* 2020;395:53–64.
10 Landewé R, Braun J, Deodhar A, *et al*. Efficacy of certolizumab pegol on signs and symptoms of axial spondyloarthritis including ankylosing spondylitis: 24-week results of a double-blind randomised placebo-controlled phase 3 study. *Ann Rheum Dis* 2014;73:39–47.
11 Rudwaleit M, van der Heijde D, Landewé R, *et al*. The development of assessment of spondyloarthritis International Society classification criteria for axial spondyloarthritis (Part II): validation and final selection. *Ann Rheum Dis* 2009;68:777–83.
12 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.
13 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.
14 Landewé R, van Tubergen A. Clinical tools to assess and monitor spondyloarthritis. *Curr Rheumatol Rep* 2015;17:47.
15 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.
16 Kiltz U, van der Heijde D, Boonen A, *et al*. Development of a health index in patients with ankylosing spondylitis (ASAS HI): final result of a global initiative based on the ICF guided by ASAS. *Ann Rheum Dis* 2015;74:830–5.
17 Ware JE, Sherbourne CD. The mos 36-item short-form health survey (SF-36). I. conceptual framework and item selection. *Med Care* 1992;30:473–83.
18 Ogdie A, Duarte-García A, Hwang M, *et al*. Measuring outcomes in axial spondyloarthritis. *Arthritis Care Res* 2020;72:47–71.
19 Jones SD, Porter J, Garrett SL, *et al*. A new scoring system for the Bath ankylosing spondylitis Metrology index (BASMI). *J Rheumatol* 1995;22:1609.
20 Qin S, Nelson L, McLeod L, *et al*. Assessing test-retest reliability of patient-reported outcome measures using intraclass correlation

coefficients: recommendations for selecting and documenting the analytical formula. *Qual Life Res* 2019;28:1029–33.
21 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
22 Navarro-Compán V, van der Heijde D, Ahmad HA, *et al*. Measurement error in the assessment of radiographic progression in rheumatoid arthritis (rA) clinical trials: the smallest detectable change (SDC) revisited. *Ann Rheum Dis* 2014;73:1067–70.
23 Auleley G-R, Benbouazza K, Spoorenberg A, *et al*. Evaluation of the smallest detectable difference in outcome or process variables in ankylosing spondylitis. *Arthritis Rheum* 2002;47:582–7.
24 Bautista-Molano W, Landewé RBM, Kiltz U, *et al*. Validation and reliability of translation of the ASAS health index in a Colombian Spanish-speaking population with spondyloarthritis. *Clin Rheumatol* 2018;37:3063–8.
25 Kiltz U, van der Heijde D, Boonen A, *et al*. Measurement properties of the ASAS health index: results of a global study in patients with axial and peripheral spondyloarthritis. *Ann Rheum Dis* 2018;77:1311–7.
26 Kiltz U, Winter J, Schirmer M, *et al*. [Validation of the German translation of the ASAS health index : A questionnaire to assess functioning and health in patients with spondyloarthritis]. *Z Rheumatol* 2019;78:352–8.
27 Kviatkovsky MJ, Ramiro S, Landewé R, *et al*. The minimum clinically important improvement and Patient-acceptable symptom state in the BASDAI and BASFI for patients with ankylosing spondylitis. *J Rheumatol* 2016;43:1680–6.
28 Kwan YH, Aw FF, Fong W, *et al*. Validity and reliability of the assessment of spondyloarthritis International Society health index in English-speaking patients with axial spondyloarthritis in Singapore. *Int J Rheum Dis* 2019;22:1644–51.
29 Machado P, Landewé R, Lie E, *et al*. Ankylosing spondylitis disease activity score (ASDAS): defining cut-off values for disease activity states and improvement scores. *Ann Rheum Dis* 2011;70:47–53.
30 Madsen OR, Rytter A, Hansen LB, *et al*. Reproducibility of the Bath ankylosing spondylitis indices of disease activity (BASDAI), functional status (BASFI) and overall well-being (BAS-G) in anti-tumour necrosis factor-treated spondyloarthropathy patients. *Clin Rheumatol* 2010;29:849–54.
31 Ozer HTE, Sarpel T, Gulek B, *et al*. Evaluation of the Turkish version of the Bath ankylosing spondylitis patient global score (BAS-G). *Clin Rheumatol* 2006;25:136–9.
32 Salaffi F, Gasparini S, Ciapetti A, *et al*. Usability of an innovative and interactive electronic system for collection of patient-reported data in axial spondyloarthritis: comparison with the traditional paper-administered format. *Rheumatology* 2013;52:2062–70.
33 Tubach F, Ravaud P, Martin-Mola E, *et al*. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: results from a prospective multina. *Arthritis Care Res* 2012;64:1699–707.
34 van Tubergen A, Black PM, Coteur G. Are patient-reported outcome instruments for ankylosing spondylitis fit for purpose for the axial spondyloarthritis patient? A qualitative and psychometric analysis. *Rheumatology* 2015;54:1842–51.
35 van Tubergen A, Coenen J, Landewé R, *et al*. Assessment of fatigue in patients with ankylosing spondylitis: a psychometric analysis. *Arthritis Rheum* 2002;47:8–16.
36 Van Tubergen A, Debats I, Ryser L, *et al*. Use of a numerical rating scale as an answer modality in ankylosing spondylitis-specific questionnaires. *Arthritis Rheum* 2002;47:242–8.
37 Garrido-Castro JL, Curbelo R, Mazzucchelli R, *et al*. High reproducibility of an automated measurement of mobility for patients with axial spondyloarthritis. *J Rheumatol* 2018;45:1383–8.
38 Haywood KL, Garratt AM, Jordan K, *et al*. Spinal mobility in ankylosing spondylitis: reliability, validity and responsiveness. *Rheumatology* 2004;43:750–7.
39 Maksymowych WP, Mallon C, Richardson R, *et al*. Development and validation of the Edmonton ankylosing spondylitis Metrology index. *Arthritis Rheum* 2006;55:575–82.
40 Marques ML, Ramiro S, Goupille P, *et al*. Measuring spinal mobility in early axial spondyloarthritis: does it matter? *Rheumatology* 2019;58:1597–606.
41 Martindale JH, Sutton CJ, Goodacre L. An exploration of the inter- and intra-rater reliability of the Bath ankylosing spondylitis Metrology index. *Clin Rheumatol* 2012;31:1627–31.
42 Ramiro S, van Tubergen A, Stolwijk C, *et al*. Reference intervals of spinal mobility measures in normal individuals: the mobility study. *Ann Rheum Dis* 2015;74:1218–24.
43 Viitanen JV, Heikkilä S, Kokko ML, *et al*. Clinical assessment of spinal mobility measurements in ankylosing spondylitis: a compact set for follow-up and trials? *Clin Rheumatol* 2000;19:131–7.
44 Frost MH, Reeve BB, Liepa AM, *et al*. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health* 2007;10 Suppl 2:S94–105.
45 Landewé RBM, van der Heijde D. Use of multidimensional composite scores in rheumatology: parsimony versus subtlety. *Ann Rheum Dis* 2020. doi:10.1136/annrheumdis-2020-216999. [Epub ahead of print: 03 Nov 2020].
46 Lee KM, Lee J, Chung CY, *et al*. Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clin Orthop Surg* 2012;4:149–55.
47 Chen M-H, Lee M-H, Liao H-T, *et al*. Health-Related quality of life outcomes in patients with rheumatoid arthritis and ankylosing spondylitis after tapering biologic treatment. *Clin Rheumatol* 2018;37:429–38.
48 Davis JC, Revicki D, van der Heijde DMF, van der Heijde DMF, *et al*. Health-Related quality of life outcomes in patients with active ankylosing spondylitis treated with adalimumab: results from a randomized controlled study. *Arthritis Rheum* 2007;57:1050–7.
49 van der Heijde D, Deodhar A, Braun J, *et al*. The effect of golimumab therapy on disease activity and health-related quality of life in patients with ankylosing spondylitis: 2-year results of the GO-RAISE trial. *J Rheumatol* 2014;41:1095–103.