RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Validation of SPARCC MRI-RETIC e-tools for increasing scoring proficiency of MRI sacroiliac joint lesions in axial spondyloarth

Walter Maksymowych ,[1,2] Anna Enevold Fløistrup E F Hadsbjerg ,[3,4] Mikkel Østergaard,[3,4] Raphael Micheroli ,[5] Susanne Juhl Pedersen ,[4,6] Adrian Ciurea ,[5] Nora Vladimirova,[3,4] Michael S Nissen ,[7] Kristyna Bubova,[8] Stephanie Wichuk,[9] Manouk de Hooge ,[10,11] Ashish J Mathew ,[12,13] Karlo Pintaric,[14] Monika Gregová,[8] Ziga Snoj,[14] Marie Wetterslev ,[3,4] Karel Gorican,[15] Burkhard Möller,[16] Iris Eshed ,[17,18] Joel Paschke,[19] Robert GW Lambert[20]

To cite: Maksymowych W, Hadsbjerg AEFEF, Østergaard M, *et al*. Validation of SPARCC MRI-RETIC e-tools for increasing scoring proficiency of MRI sacroiliac joint lesions in axial spondyloarth. *RMD Open* 2024;**10**:e003923. doi:10.1136/rmdopen-2023-003923

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/rmdopen-2023-003923).

Received 19 November 2023
Accepted 19 January 2024

For numbered affiliations see end of article.

Correspondence to
Dr Walter Maksymowych; WALTER.MAKSYMOWYCH@UALBERTA.CA

ABSTRACT

Background The Spondyloarthritis Research Consortium of Canada (SPARCC) developers have created web-based calibration modules for the SPARCC MRI sacroiliac joint (SIJ) scoring methods. We aimed to test the impact of applying these e-modules on the feasibility and reliability of these methods.

Methods The SPARCC-SIJ$_{RETIC}$ e-modules contain cases with baseline and follow-up scans and an online scoring interface. Visual real-time feedback regarding concordance/discordance of scoring with expert readers is provided by a colour-coding scheme. Reliability is assessed in real time by intraclass correlation coefficient (ICC), cases being scored until ICC targets are attained. Participating readers (n=17) from the EuroSpA Imaging project were randomised to one of two reader calibration strategies that each comprised three stages. Baseline and follow-up scans from 25 cases were scored after each stage was completed. Reliability was compared with a SPARCC developer, and the System Usability Scale (SUS) assessed feasibility.

Results The reliability of readers for scoring bone marrow oedema was high after the first stage of calibration, and only minor improvement was noted following the use of the inflammation module. Greater enhancement of reader reliability was evident after the use of the structural module and was most consistently evident for the scoring of erosion (ICC status/change: stage 1 (0.42/0.20) to stage 3 (0.50/0.38)) and backfill (ICC status/change: stage 1 (0.51/0.19) to stage 3 (0.69/0.41)). The feasibility of both e-modules was evident by high SUS scores.

Conclusion The SPARCC-SIJ$_{RETIC}$ e-modules are feasible, effective knowledge transfer tools, and their use is recommended before using the SPARCC methods for clinical research and tria

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Objective assessment of inflammatory and structural lesions on MRI of the sacroiliac joint (SIJ) in axial spondyloarthritis clinical trials and research can be done effectively using the Spondyloarthritis Research Consortium of Canada (SPARCC) MRI SIJ scoring methods, which are instruments that are now included in the Assessments in SpondyloArthritis International Society core set.

WHAT THIS STUDY ADDS

⇒ The SPARCC developers created two interactive web-based knowledge transfer (KT) e-modules, which reflect the scoring rules set by the developers and permit training and ongoing calibration of successive generations of readers, which were validated per Outcome Measures in Rheumatology (OMERACT) recommendations for enhancing scoring proficiency of untrained and even trained readers in the use of the SPARCC methods.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These SPARCC e-modules provide a template for the development and validation of KT tools for imaging-based scoring instruments that are considered essential in the OMERACT framework for the routine calibration of readers prior to the use of these methods in clinical research and clinical trials.

INTRODUCTION

The advent of MRI for the evaluation of axial spondyloarthritis (axSpA) marks a milestone not only for enhanced diagnostic accuracy but also for disease classification.[1] MRI inflammation has also been used as an endpoint in

randomised placebo-controlled trials (RCTs) of biological disease-modifying antirheumatic drugs (DMARDs) in axSpA and, more recently, in RCTs of targeted synthetic DMARDs.[2–20] Scoring methodologies, such as the Berlin and Spondyloarthritis Research Consortium of Canada (SPARCC) methods, are based on semiquantitative assessment of MRI inflammation in the sacroiliac joint (SIJ) and spine.[21]

Feasibility, reliability and discriminatory properties of these instruments according to the Outcome Measures in Rheumatology (OMERACT) filter have demonstrated their high degree of reliability and substantial capacity to discriminate between active therapy and placebo within the typical 12–16-week timeframe of placebo-controlled RCTs.[4–7 22] Moreover, an extensive analysis of the metric properties of these instruments conducted as part of a recent update of the Assessments in SpondyloArthritis International Society (ASAS) core outcome set led to the recommendation that the use of the SPARCC SIJ and spine instruments be mandatory in at least one pivotal RCT of DMARD.[23] SPARCC investigators have also developed an instrument to assess structural lesions in the SIJ and demonstrated that this instrument could also demonstrate significant differences in the extent of structural damage between active therapy and placebo within the 12–16-week timeframe of a placebo-controlled trial.[10 24–27] ASAS has endorsed this instrument as an objective tool for assessing structural lesions in RCTs of axSpA.[23]

A limitation of imaging-based scoring instruments that affects their widespread application in a manner that ensures reliable and accurate data is the lack of feasible knowledge transfer tools (KT tools). Developers have often provided published atlases with examples of images and appropriate scoring of lesions in addition to the original descriptions of these instruments. However, such publications provide only a small sample of the potential variation in imaging abnormalities, and such KT tools are not based on Digital Imaging and Communications in Medicine (DICOM) images, which would be preferable for optimal visualisation of consecutive images. Consequently, training in using such instruments has continued to entail the traditional in-person review at workstations and displays followed by iterative training exercises to ensure sufficient reliability with developer scores and data entry on Excel spreadsheets. These standard practices are time-consuming, require the availability of expert readers on site, are prone to data entry errors and do not provide legacy tools that accurately reflect the rules set by the developers and permit training and ongoing calibration of successive generations of readers even in remote settings.

The developers of the SPARCC MRI scoring methods have created two calibration modules for assessing inflammatory and structural MRI lesions in the SIJ based on consensus scores from these instrument developers and real-time iterative feedback built into an online scoring schematic that is integrated directly with the MRI image. The modules permit remote web-based training and calibration of readers with case-based imaging content in DICOM format aimed at precision in the understanding of the scoring methodology, illustration of diverse examples of inflammation and structural change on MRI scans of the SIJ, and attainment of prespecified performance targets for reader reliability. In this report, we describe the results of validation exercises aimed at testing the impact of applying these modules in the calibration process on feasibility and interobserver reliability of the SPARCC SIJ methods in multiple readers with expertise ranging from none to extensive in the prior use of these methods.

METHODS

Development of SPARCC MRI sacroiliac joint RETIC modules

The scoring of MRI lesions in the SIJ using the SPARCC methods is based on the subdivision of individual semicoronal MRI slices through the SIJ into quadrants (bone marrow oedema (BME), erosion and fat lesion) and halves (backfill and ankylosis). The two calibration systems for inflammatory and structural MRI lesions, respectively, are each comprised of (1) a PowerPoint module, which describes each scoring method in detail and provides numerous examples of images that the developers have scored and (2) a web-based interactive Real Time Iterative Calibration (RETIC) calibration module for scoring of lesions seen on MRI scans of cases with axSpA (available at www.carearthritis.com). For the latter, the presence or absence of lesions in each SIJ quadrant (BME, erosion and fat metaplasia) or half (backfill and ankylosis) is recorded dichotomously by direct online data entry using a mouse click on a web-based interface that includes a schematic of these joints adjacent to the DICOM image (figure 1, www.carearthritis.com/service/mri-scoring-modules). The interface includes individual schematic figures for each lesion, with the SIJ, divided into either quadrants or halves.

The SPARCC-SIJ$_{RETIC\text{-}INF}$ module is comprised of 50 DICOM cases, each with scans from baseline and 12 weeks after the start of tumour necrosis factor inhibitor (TNFi) therapy. The SPARCC-SIJ$_{RETIC\text{-}STR}$ module is also comprised of 50 DICOM cases, but each case includes scans from baseline and 2 years after the start of TNFi therapy. Pairs of scans from baseline and follow-up have been scored by the SPARCC developers blinded to time point by entering 0 (denoting lesion is absent) or 1 (denoting lesion is present) in fields on the SIJ quadrants or halves of the SIJ schematic. All the cases have been scored on consecutive semicoronal slices through the SIJ and discrepancies resolved by consensus at the level of each individual SIJ quadrant or half. When readers use these modules to gain familiarity with these SPARCC methods, continuous visual real-time feedback is provided regarding concordance/discordance of scoring per SIJ quadrant or half with developer scores according to a colour-coding scheme. For instance, a blue colour at the SIJ quadrant/half indicates concordance, while a red colour indicates

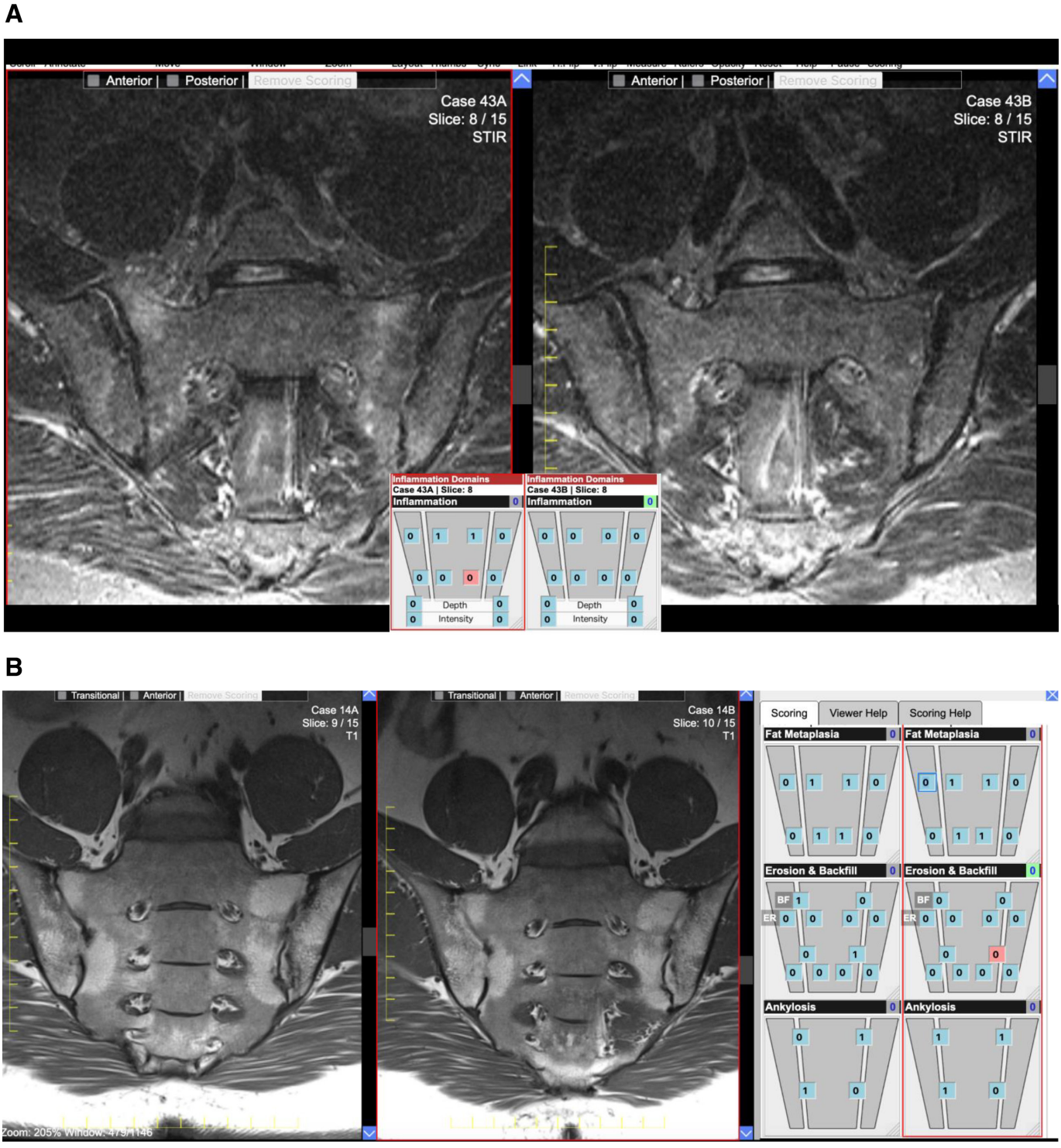

Figure 1 Spondyloarthritis Research Consortium of Canada (SPARCC) MRI sacroiliac joint calibration modules (SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$). (A) For the assessment of inflammation, bone marrow oedema (BME) is recorded as present or absent in each sacroiliac joint quadrant on a schematic, blue and red coding per quadrant denoting concordance and discordance, respectively, with SPARCC developer assignments. (B) For the assessment of structural lesions, fat metaplasia and erosion are recorded as present or absent in SIJ quadrants, while backfill and ankylosis are recorded in SIJ halves on a similar schematic and coding as BME. Demos of these modules can be accessed at www.carearthritis.com/service/mri-scoring-modules/.

discordance (figure 1). Reliability is additionally assessed in real time by the module software using the intraclass correlation coefficient (ICC), the first ICC data being provided after 10 cases. Additional ICC data are provided after successive batches of 10 cases have been scored. Accreditation for SPARCC MRI SIJ inflammation score is achieved with status and change score ICC of ≥0.8 and ≥0.7, respectively, and is based on the scoring of at least

20 cases. To be accredited as a SPARCC MRI SIJ structural score reader, the ICC attained must meet the following thresholds: fat and ankylosis status (baseline scan) score ICC ≥0.7, erosion and backfill status (baseline scan) score ICC ≥0.5 and change from baseline to follow-up score (all domains) ICC ≥0.5.

ICC targets required for structural lesions are lower than for inflammation (BME) because the amount of change between patients after the use of TNFi is much larger for BME than for structural lesions. The ICC is a relative measure of reliability that calculates the proportion of the total variance that is due to the variance between cases. Consequently, the small degree of variation in the amount of structural change between cases biases ICC score towards lower values even when interobserver reliability may be high.

Study design and reading exercises

Readers comprised 11 rheumatologists, 5 radiologists and 1 research associate, all participating in the EuroSpA Research Collaboration Network. Their reading experience was as follows, based on a questionnaire: six readers (rheumatologist n=2 and radiologist n=4) had no prior experience in reading scans with either of the SPARCC methods and minimal knowledge of the methodology, six readers (rheumatologist n=6 and radiologist n=0) were considered to have intermediate expertise based on awareness of the methodology and 1–2 scoring exercises and five (rheumatologist n=3, radiologist n=1 and research associate n=1) were considered as being experienced readers with these methods having participated in greater than or equal to six reading exercises. Readers were randomised into two groups (A and B) matched on the level of experience and educational background.

We aimed to test the performance of the SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ modules in enhancing the scoring proficiency of EuroSpA readers in comparison with SPARCC developer gold standard scores by randomising readers into one of two calibration strategies, stratified by the level of experience and educational background. The exercise consisted of 3 calibration activities and the scoring of 3 different image sets of 25 cases after each step of calibration in both strategies and separately for each scoring method so that 75 cases in total were scored for SPARCC inflammation and 75 different cases for SPARCC structural (figure 2). None of these 75 cases are replicated in the RETIC scoring modules, each of which contains 50 entirely separate cases.

Each case had baseline and follow-up scans, and readers were blinded to the chronology of the scans. In both strategies, all readers first reviewed the original manuscript describing the methodology of the SPARCC MRI SIJ inflammation method, then scored 25 cases using this method and then reviewed the original manuscript describing the methodology of the SPARCC MRI SIJ structural scoring method followed by the scoring of 25 different cases using this method. Subsequent calibration activities were as follows:

A. In strategy A (readers in group A), step 2 consisted of readers reviewing PowerPoint instructions for the SPARCC inflammation method as well as the use of the web-based SPARCC-SIJ$_{RETIC\text{-}INF}$ module and then the scoring of 25 cases using this method. This was

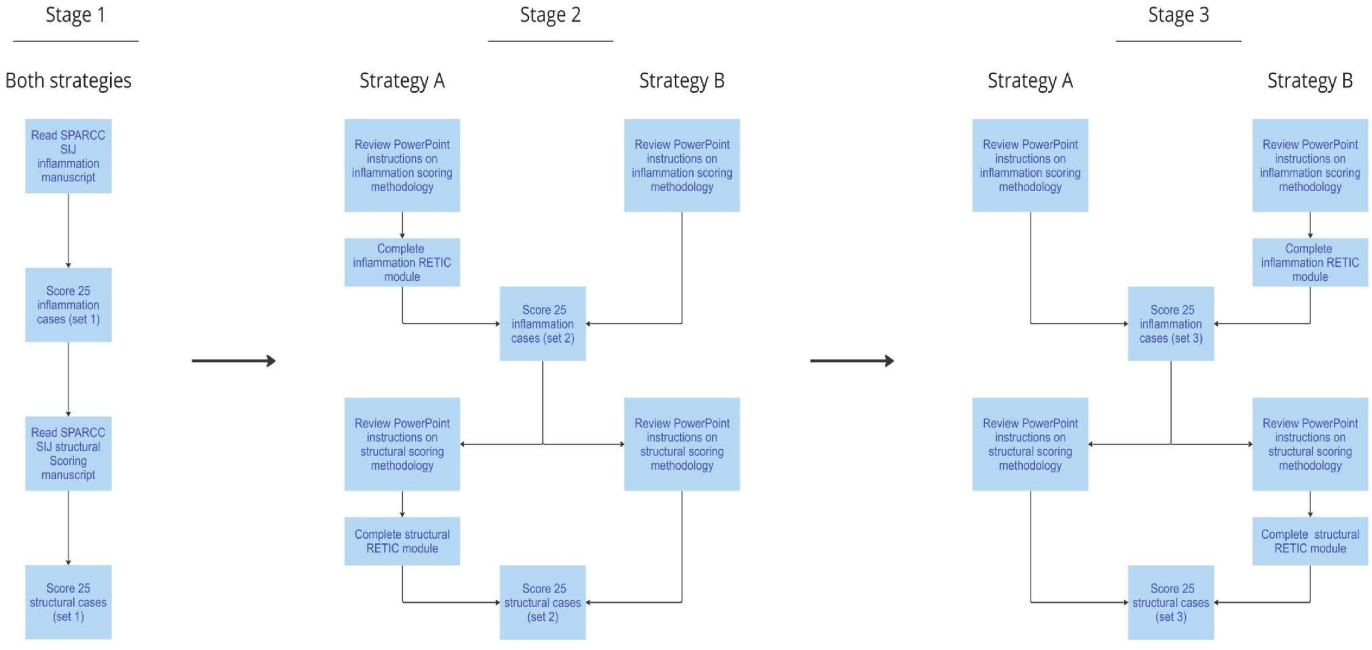

Figure 2 Calibration activities and reading exercises for two strategies of calibration to test the impact of the Spondyloarthritis Research Consortium of Canada (SPARCC) RETIC modules in enhancing scoring proficiency when using the SPARCC MRI sacroiliac joint inflammation and structural scoring methods. The RETIC modules require scoring of between 20 and 50 MRI cases and achievement of scoring proficiency according to intraclass correlation coefficient targets for status and change scores specified on www.carearthritis.com. SPARCC SIJ, Spondyloarthritis Research Consortium of Canada sacroiliac joint.

followed by a review of the PowerPoint instructions for the SPARCC structural method as well as the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module and then the scoring of 25 different cases using this method. In the third and final step, readers rereviewed the PowerPoint instructions for SPARCC inflammation and then scored 25 cases using this method, followed by a rereview of the PowerPoint instructions for SPARCC structural and then the scoring of 25 different cases using this method.

B. In strategy B (readers in group B), step 2 consisted of readers only reviewing the PowerPoint instructions for the SPARCC inflammation method, then the scoring of 25 cases using this method, followed by a review of PowerPoint instructions for the SPARCC structural method and the scoring of 25 different cases using this method. In the third and final step, readers rereviewed PowerPoint instructions for SPARCC inflammation but then also used the SPARCC-SIJ$_{RETIC\text{-}INF}$ module before scoring the final 25 cases with this method. This was followed by a rereview of PowerPoint instructions for SPARCC structural method as well as the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module before scoring 25 cases with this method.

When scoring inflammation, both T1-weighted and Short Tau Inversion Recovery (STIR) images were available, while when scoring structural changes, only T1-weighted images were available. All the test cases had previously been scored by the developers. Selection of these cases for each of the three calibration steps was aimed at a comparable level of disease severity for each set of 25 cases as determined by developer mean SPARCC scores for inflammatory and structural lesions. This was desirable so that differences in reliability from one reading exercise to the next could be reasonably ascribed to the calibration activity rather than differences in the degree of difficulty in scoring the MRI scans.

Assessment of feasibility

The feasibility of using the RETIC calibration modules as well as the SPARCC methods was assessed by recording the time expended on the reading of each case, which was done automatically by the reading software, and by completing the System Usability Scale (SUS)[28] (www.usability.gov). SUS is a simple, 10-item attitude Likert scale giving a global view of subjective assessments of usability. It yields a single score on a scale of 0–100, with higher scores indicating higher perceived usability.[29] This scale has been widely used in evaluating a range of systems and has led to normative data so that raw SUS scores can be converted into percentile ranks.[30] The 50th percentile score is 68 and is generally regarded as the cut-off for an instrument likely to be widely applied. EuroSpA readers were asked to rate each SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ module using SUS after completion of each module and also to rate each SPARCC scoring method after having completed the entire reading exercise.

Statistics

Frequencies of each SIJ lesion were assessed descriptively. The reliability for the number of SIJ quadrants or halves with SIJ lesions was assessed by ICC 2.1 (two-way random effects, absolute agreement and single rater/measurement MedCalc V.12.6) for each of the three reading exercises. We assessed interobserver reliability in a pairwise manner by comparing each reader's scores with a SPARCC developer Musculoskeletal radiologist (RL). Mean (SD) ICC scores were calculated, and the results are presented according to the calibration strategy (Group A or B) and also according to the prior level of reader expertise with these methods (none to extensive).

RESULTS

Study populations and calibration activities

Baseline demographics were typical of patients diagnosed with axSpA and meeting modified New York classification criteria for each of the 3 sets of 25 cases whose baseline and follow-up SIJ MRI scans were evaluated using the SPARCC methods. The majority were human leucocyte antigen B27-positive males starting a TNFi with mean symptom duration greater than 10 years (Online supplemental table 1). For the SPARCC-SIJ$_{RETIC\text{-}INF}$ module, all readers achieved prespecified target ICCs for BME (≥0.70 for change score and ≥0.80 for status score). When using the RETIC module, the average number of cases that had to be scored for BME in order to reach the prespecified target using the SPARCC SIJ inflammation score was 31 (range 20–50). For the SPARCC-SIJ$_{RETIC\text{-}STR}$ module, all readers achieved prespecified target ICCs for ankylosis (≥0.50 for change score and ≥0.70 for status score) and backfill (≥0.50 for change and status score). One reader did not achieve the prespecified target ICC for erosion (≥0.50 for change and status score; reader ICC change score for erosion=0.47), and one reader did not achieve the prespecified target ICC for fat lesion (≥0.50 for change score and ≥0.70 for status score; reader ICC status score for fat lesion=0.56). The average number of cases that had to be scored to reach prespecified targets for structural lesions using the SPARCC SIJ structural score was 45 (range 20–90) (fat lesion, 20 (range: 20–20) (excludes the reader who did not achieve the ICC target); erosion, 42 (range: 20–90); backfill, 22 (range: 20–40); and ankylosis, 21 (range: 20–30)).

MRI characteristics of the study populations

SPARCC developer scores for inflammatory and structural MRI lesions were comparable between the 3 sets of 25 cases with paired baseline and follow-up MRI scans. There was a much greater change between baseline and follow-up scans in BME than for structural lesions (table 1). There were no significant differences in status or baseline to follow-up change scores for BME between the three sets of cases and between these three sets of

Table 1 Baseline imaging characteristics for three sets of baseline and follow-up MRI scans (n=25 each) from patients with axSpA that were assessed for BME and structural lesions in the SIJ in three reading exercises by the two SPARCC method developers

	Stage 1 cases			Stage 2 cases			Stage 3 cases		
MRI feature	Baseline	Follow-up	ICC baseline and change	Baseline	Follow-up	ICC baseline and change	Baseline	Follow-up	ICC baseline and change
SPARCC BME*									
Reader 1	7.8 (9.0)	2.7 (6.9)	BL: 0.88	10.9 (13.1)	0.8 (1.9)	BL: 0.93	12.4 (15.4)	3.4 (4.5)	BL: 0.93
Reader 2	9.0 (10.9)	3.0 (7.0)	Change: 0.94	13.0 (15.8)	1.8 (4.5)	Change: 0.94	14.0 (14.2)	5.2 (10.3)	Change: 0.90
SPARCC SSS erosion*									
Reader 1	4.7 (5.1)	2.0 (2.8)	BL: 0.62	4.5 (5.3)	1.8 (3.1)	BL: 0.57	3.8 (4.1)	2.3 (3.8)	BL: 0.67
Reader 2	7.6 (5.4)	4.2 (4.2)	Change: 0.59	9.5 (7.0)	5.6 (4.0)	Change: 0.63	7.1 (5.4)	4.5 (4.4)	Change: 0.75
SPARCC SSS fat*									
Reader 1	5.0 (7.8)	5.5 (8.1)	BL: 0.87	4.3 (5.8)	5.4 (6.0)	BL: 0.90	6.6 (8.4)	8.2 (9.1)	BL: 0.93
Reader 2	4.7 (7.1)	6.0 (8.3)	Change: 0.76	5.8 (6.9)	7.8 (8.5)	Change: 0.74	6.7 (8.7)	8.3 (9.2)	Change: 0.73
SPARCC SSS backfill*									
Reader 1	1.1 (2.3)	1.6 (2.5)	BL: 0.83	1.0 (2.1)	0.9 (1.4)	BL: 0.67	1.6 (3.5)	1.9 (3.7)	BL: 0.75
Reader 2	1.7 (3.4)	2.1 (3.4)	Change: 0.40	2.0 (2.3)	2.1 (2.6)	Change: 0.15	2.8 (4.2)	3.0 (4.0)	Change: 0.34
SPARCC SSS ankylosis*									
Reader 1	2.8 (5.4)	3.2 (5.9)	BL: 0.97	1.3 (2.8)	1.8 (3.8)	BL: 0.87	1.3 (3.5)	1.9 (3.7)	BL: 0.95
Reader 2	3.5 (6.4)	3.8 (7.3)	Change: 0.80	2.0 (3.5)	3.0 (4.7)	Change: 0.76	1.7 (3.8)	3.0 (5.1)	Change: 0.72

*Spondyloarthritis Research Consortium of Canada developer scores (mean (SD)) for baseline and follow-up time points.
BL, baseline; BME, bone marrow oedema; ICC, intraclass correlation coefficient; SPARCC, Spondyloarthritis Research Consortium of Canada; SSS, Sacroiliac Joint Structural Score.

cases and the cases in the SPARCC-SIJ$_{RETIC\text{-}INF}$ module. For structural lesions, there were also no significant differences in status or change scores between the three sets of cases, but comparisons of cases in the SPARCC-SIJ$_{RETIC\text{-}STR}$ module indicated significantly lower scores for erosion at baseline from cases in stages 2 and 3 and significantly lower scores for backfill at baseline from cases in stage 2 (data not shown). Among structural lesions, scores for erosion decreased, while scores for fat lesions, backfill and ankylosis increased from baseline to follow-up. The degree of change for structural lesions was highest for erosion and lowest for backfill, especially for backfill scores in stage 2 scans, where the mean change for one developer reader was a decrease in score of 0.1, while the mean change for the second developer reader was an increased score of 0.1. Interobserver reliability for baseline and change scores between SPARCC developers were similar between the 3 sets of 25 cases, being much higher for BME than for structural lesions, commensurate with the much lower degree of change for structural than BME lesions (table 1). This was particularly evident on the reliability of the assessment of change in backfill, especially for the 25 cases assessed at stage 2, which was much worse than for the 25 cases assessed at stages 1 and 3.

MRI readings by EuroSpA readers

Reliability/SPARCC MRI SIJ inflammation scores

The reliability of EuroSpA readers with the SPARCC developer radiologist for scoring the extent of BME on baseline MRI scans and detecting change in degree of BME from baseline to follow-up was high (≥0.80) even after stage 1 of the reading exercise, irrespective of the prior experience of the readers (table 2). Moreover, reliability was almost comparable with the reliability noted between the two SPARCC developers scoring the same cases (table 1). There was no consistent effect of applying the SPARCC-SIJ$_{RETIC\text{-}INF}$ module in strategy A (between reading cases at stages 1 and 2). Although an effect of the module was apparent to a minor degree for strategy B (between reading cases at stages 2 and 3), especially for status scores and in the least experienced readers (table 2, figure 3), there were no consistent differences between the strategies in reliability attained by the completion of calibration and after reading stage 3 cases (table 2).

Reliability/SPARCC MRI SIJ Structural Scores

The reliability of EuroSpA readers with the SPARCC developer radiologist for scoring the extent of erosions on baseline MRI scans and change in degree of erosion from baseline to follow-up was lower than for BME, commensurate with the lesser degree of change in this structural outcome and the morphological complexity of these lesions. By the completion of the entire exercise, experienced EuroSpA readers were approaching similar reliability to that noted between the two SPARCC developers for baseline erosion scores but much less so for detecting a change in erosion (tables 1 and 2, online supplemental table 2). A significant increase in EuroSpA reader reliability was noted after using the SPARCC-SIJ$_{RETIC\text{-}STR}$ module for detecting the extent of erosion at baseline and also for detecting a change in erosion, irrespective of reader expertise or strategy for calibration, which was most evident for the least experienced readers (figure 4A). By the completion of all calibration activities and after assessment of stage 3 cases, the reliability for both baseline and change scores had improved compared with stage 1 and was comparable among readers irrespective of strategy or prior experience of the readers (table 2, figure 5).

Similar observations were noted for the assessment of backfill and fat lesions by EuroSpA readers. However, more consistent enhancement of reader reliability after use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was found for strategy B (figure 4B,C). By the completion of calibration activities, the reliability for the assessment of backfill and fat had improved and was comparable among readers irrespective of strategy or prior experience of the readers (table 2, figure 5). The only exception was a decrease in the reliability for change in fat lesion scores among readers of intermediate expertise who were randomised to strategy B.

For the reliable detection of ankylosis, enhanced reliability after the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was only noted for readers randomised to strategy B, while deterioration in reader reliability after the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was noted for strategy A (figure 4D). However, the reliability for both baseline and change scores in ankylosis had improved by the completion of all calibration activities after the assessment of stage 3 cases compared with stage 1 cases and was comparable among readers irrespective of strategy or prior experience of the readers (table 2, figure 5). It should be noted that reliability between SPARCC developers for ankylosis and backfill was worse for stage 2 cases when compared with either stage 1 or stage 3 cases (table 1).

It is noteworthy that some individual pairs of readers achieved reliability comparable with the SPARCC developers irrespective of prior experience with the SPARCC structural method (example provided in online supplemental figure 1).

Feasibility

SPARCC-SIJ$_{RETIC\text{-}INF}$ module

The mean time expended by SPARCC developers for the paired evaluation of baseline and follow-up scans of each individual case for BME was 5–6 min at each of stages 1–3 (online supplemental table 3), while the mean time per EuroSpA reader decreased from 8 min for stage 1 cases to 5.4 min for stage 3 cases. For EuroSpA readers randomised to strategy A, the mean time decreased from 7.9 min at stage 1 to 6.4 min at stage 2, following the use of the SPARCC SIJ inflammation RETIC calibration module. For EuroSpA readers randomised to strategy B, the mean time was 8.1 min at stage 1 and then decreased from 8.2 min at stage 2 to 5.7 min at stage 3, following the

Table 2 Mean ICC (compared with radiologist SPARCC developer) for status/change scores in BME and structural lesions in the SIJ according to the calibration strategy

		Strategy A* Baseline ICC (95% CI) Change ICC {95% CI]				Strategy B* Baseline ICC (95% CI) Change ICC {95% CI]				All readers† Baseline ICC (95% CI) Change ICC (95% CI)	
MRI lesion	**Readers, N**	**Stage 1 cases (n=25)**	**Stage 2 cases (n=25)**	**Stage 3 cases (n=25)**	**Readers, N**	**Stage 1 cases (n=25)**	**Stage 2 cases (n=25)**	**Stage 3 cases (n=25)**	**Readers, N**	**Stage 1 cases (n=25)**	**Stage 3 cases (n=25)**
BME	**8**	0.89 (0.86 to 0.93) 0.91 (0.88 to 0.94)	0.89 (0.86 to 0.92) 0.88 (0.84 to 0.91)	0.89 (0.83 to 0.94) 0.88 (0.83 to 0.93)	**9**	0.84 (0.77 to 0.91) 0.90 (0.86 to 0.94)	0.80 (0.73 to 0.88) 0.85 (0.80 to 0.90)	0.89 (0.85 to 0.92) 0.86 (0.82 to 0.90)	**17**	0.87 (0.82 to 0.91) 0.90 (0.88 to 0.93)	0.89 (0.86 to 0.92) 0.87 (0.84 to 0.90)
Backfill	**8**	0.39 (0.11 to 0.67) 0.17 (0.02 to 0.32)	0.57 (0.50 to 0.64) 0.35 (0.23 to 0.47)	0.72 (0.60 to 0.85) 0.46 (0.34 to 0.58)	**9**	0.61 (0.44 to 0.78) 0.21 (0.08 to 0.33)	0.29 (0.19 to 0.39) 0.18 (0.03 to 0.32)	0.66 (0.55 to 0.77) 0.36 (0.22 to 0.50)	**17**	0.51 (0.34 to 0.67) 0.19 (0.10 to 0.28)	0.69 (0.61 to 0.77) 0.41 (0.31 to 0.50)
Erosion	**8**	0.39 (0.21 to 0.56) 0.11 (-0.02 to 0.25)	0.60 (0.51 to 0.69) 0.24 (0.11 to 0.37)	0.50 (0.42 to 0.58) 0.39 (0.29 to 0.49)	**9**	0.44 (0.32 to 0.57) 0.28 (0.18 to 0.38)	0.42 (0.29 to 0.55) 0.14 (0.07 to 0.21)	0.50 (0.40 to 0.60) 0.36 (0.24 to 0.49)	**17**	0.42 (0.31 to 0.52) 0.20 (0.11 to 0.29)	0.50 (0.44 to 0.56) 0.38 (0.30 to 0.46)
Fat lesion	**8**	0.57 (0.42 to 0.72) 0.45 (0.32 to 0.59)	0.60 (0.43 to 0.76) 0.21 (0.02 to 0.4)	0.78 (0.63 to 0.92) 0.52 (0.37 to 0.67)	**9**	0.52 (0.40 to 0.64) 0.56 (0.44 to 0.68)	0.52 (0.36 to 0.68) 0.16 (-0.08 to 0.39)	0.82 (0.75 to 0.88) 0.49 (0.36 to 0.62)	**17**	0.54 (0.45 to 0.64) 0.51 (0.42 to 0.60)	0.80 (0.72 to 0.87) 0.51 (0.41 to 0.60)
Ankylosis	**8**	0.86 (0.79 to 0.93) 0.67 (0.54 to 0.79)	0.78 (0.72 to 0.84) 0.39 (0.22 to 0.56)	0.87 (0.82 to 0.92) 0.80 (0.75 to 0.84)	**9**	0.84 (0.73 to 0.95) 0.49 (0.27 to 0.72)	0.64 (0.54 to 0.74) 0.39 (0.27 to 0.52)	0.91 (0.87 to 0.95) 0.80 (0.77 to 0.84)	**17**	0.85 (0.79 to 0.92) 0.57 (0.44 to 0.71)	0.89 (0.86 to 0.93) 0.80 (0.77 to 0.83)

*For strategy A readers, the RETIC module was applied between stages 1 and 2, while in strategy B readers, it was done between stages 2 and 3.
†All readers scored stage 1 cases after a review of the Spondyloarthritis Research Consortium of Canada manuscripts, and all readers had conducted a review of both the PowerPoint and RETIC modules before the scoring of stage 3 cases.
BME, bone marrow oedema; ICC, intraclass correlation coefficient.

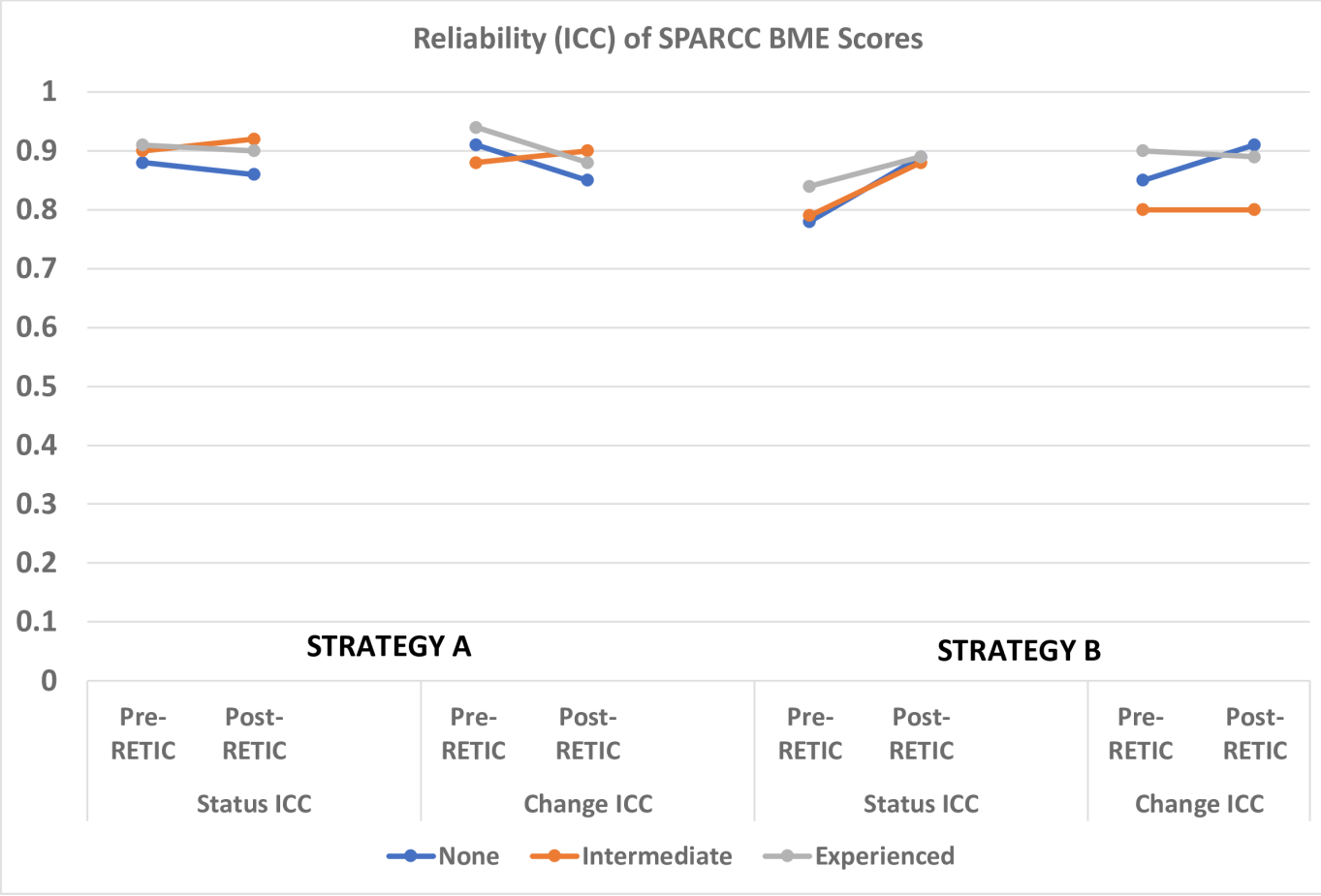

Figure 3 Reliability (mean intraclass correlation coefficient) between EuroSpA readers, stratified by the level of prior expertise with Spondyloarthritis Research Consortium of Canada (SPARCC) methods, and SPARCC developer radiologist for baseline and change in SPARCC bone marrow oedema (BME) scores prior to and after EuroSpA reader calibration using the SPARCC-SIJ$_{RETIC\text{-}INF}$ module. The RETIC module was applied for strategy A readers between stages 1 and 2, while it was performed between stages 2 and 3 in strategy B readers. BME, bone marrow oedema; ICC, intraclass correlation coefficient; SPARCC, Spondyloarthritis Research Consortium of Canada.

use of the SPARCC-SIJ$_{RETIC\text{-}INF}$ module. By the completion of the exercise, the mean time expended by EuroSpA readers was comparable with SPARCC developers. The mean (SD) (range) SUS score for the SPARCC-SIJ$_{RETIC\text{-}INF}$ module was 76.0 (14.4) (42.5–95), and for the SPARCC SIJ inflammation method, the mean score was 76.8 (14.4) (45–100). The scores for each reader are provided in online supplemental table 4.

SPARCC-SIJ$_{RETIC\text{-}STR}$ module

The mean time expended by SPARCC developers for the paired evaluation of baseline and follow-up scans of each individual case for structural lesions was 9.2 min for stage 1, 13.1 min for stage 2 and 11.8 min for stage 3 (online supplemental table 3). The mean time per EuroSpA reader was 9.9 min for stage 1, 9.2 min for stage 2 and 7.6 min for stage 3 cases. For EuroSpA readers randomised to strategy A, the mean time increased from 10 min at stage 1 to 10.4 min at stage 2, following the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module. For EuroSpA readers randomised to strategy B, the mean time was 9.8 min at stage 1 and then decreased from 8.2 min at stage 2 to 7.5 min at stage 3, following the use of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module. The mean (SD) (range) SUS score for the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was 71.0 (15.9) (27.5–95), and for the SPARCC SIJ structural method, the mean score was 74.0 (16.9) (30–100).

SUS scores for both SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ modules were ≥68 for the majority of readers (76.5% and 70.6% for the inflammation and structural modules, respectively). However, this was more frequently observed for intermediate and experienced readers (figure 6).

DISCUSSION

We have developed novel web-based calibration modules for the SPARCC MRI SIJ inflammation and structural scoring methods based on DICOM images, real-time iterative feedback and prespecified targets for attaining scoring proficiency, which have been validated in this multireader exercise that included readers with varying levels of expertise with the SPARCC scoring methods.

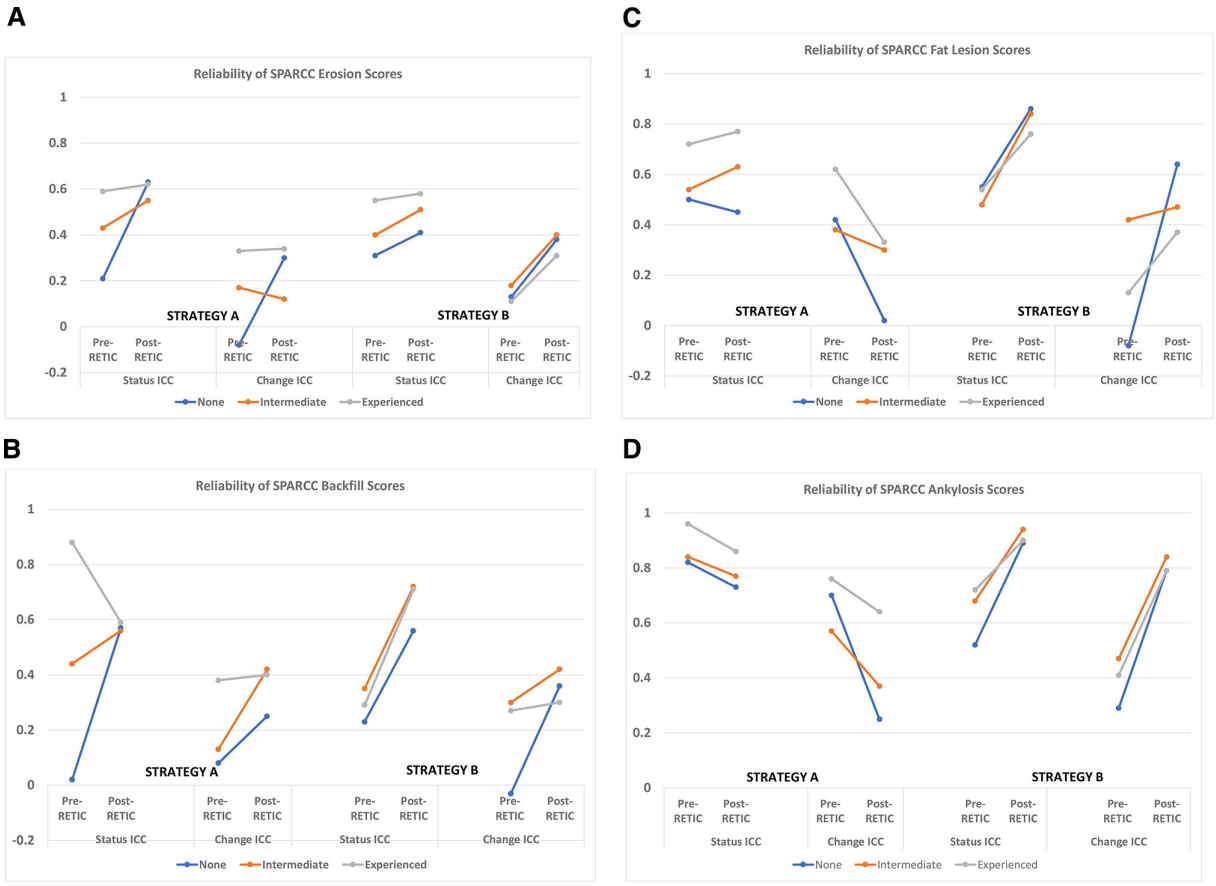

Figure 4 Reliability (mean intraclass correlation coefficient) between EuroSpA readers, stratified by the level of prior expertise with Spondyloarthritis Research Consortium of Canada (SPARCC) methods, and SPARCC developer radiologist for baseline and change in SPARCC structural scores prior to and after EuroSpA reader calibration using the SPARCC-SIJ$_{RETIC\text{-}STR}$ module. For strategy A readers, the RETIC module was applied between stages 1 and 2, while in strategy B readers, it was performed between stages 2 and 3.

The SPARCC SIJ inflammation scoring method was readily understood and adopted, including by inexperienced readers, as demonstrated by the high values attained for interobserver reliability with the SPARCC developer radiologist, comparable with the reliability between the two SPARCC method developers. Furthermore, incremental gains in reader reliability after the use of the SPARCC-SIJ$_{RETIC\text{-}INF}$ module were relatively minor. Conversely, a much greater enhancement of reader reliability was evident for the SPARCC SIJ structural damage scoring method after the SPARCC-SIJ$_{RETIC\text{-}STR}$ module, and this was greatest for inexperienced readers and most consistently evident for the scoring of erosion and backfill. The outcomes were less clear for the scoring of fat lesions and ankylosis. However, the reliability for structural lesion scores had improved by the completion of all calibration activities and was comparable among readers irrespective of strategy or prior experience of the readers. Moreover, some individual pairs of readers achieved reliability comparable with the SPARCC developers, irrespective of prior experience with the SPARCC structural method, documenting that further reader proficiency can be achieved with further training.

Both SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ modules and the scoring methods were considered feasible as judged by the reading times to score each case, which were comparable with SPARCC developer times, and the high SUS scores from the majority of readers, which were above the cut-off for an instrument likely to be widely applied based on extensive experience with this instrument.[30]

Recent consensus-based deliberations conducted by imaging and methodology experts of the OMERACT consortium have resulted in the drafting of a framework of recommendations aimed at reducing the sources of variability for imaging-based instruments.[31] Moreover, it was considered essential that these be implemented in operational guidelines for the application of an imaging instrument because reader reliability, especially for detecting change, influences responsiveness and the ability of an instrument to discriminate between therapeutic interventions. The recommendations stipulated the importance of a clear description of the scoring framework, the availability of reference standards such as an atlas of images and a systematic process for training using validated KT tools. These OMERACT recommendations also stipulate

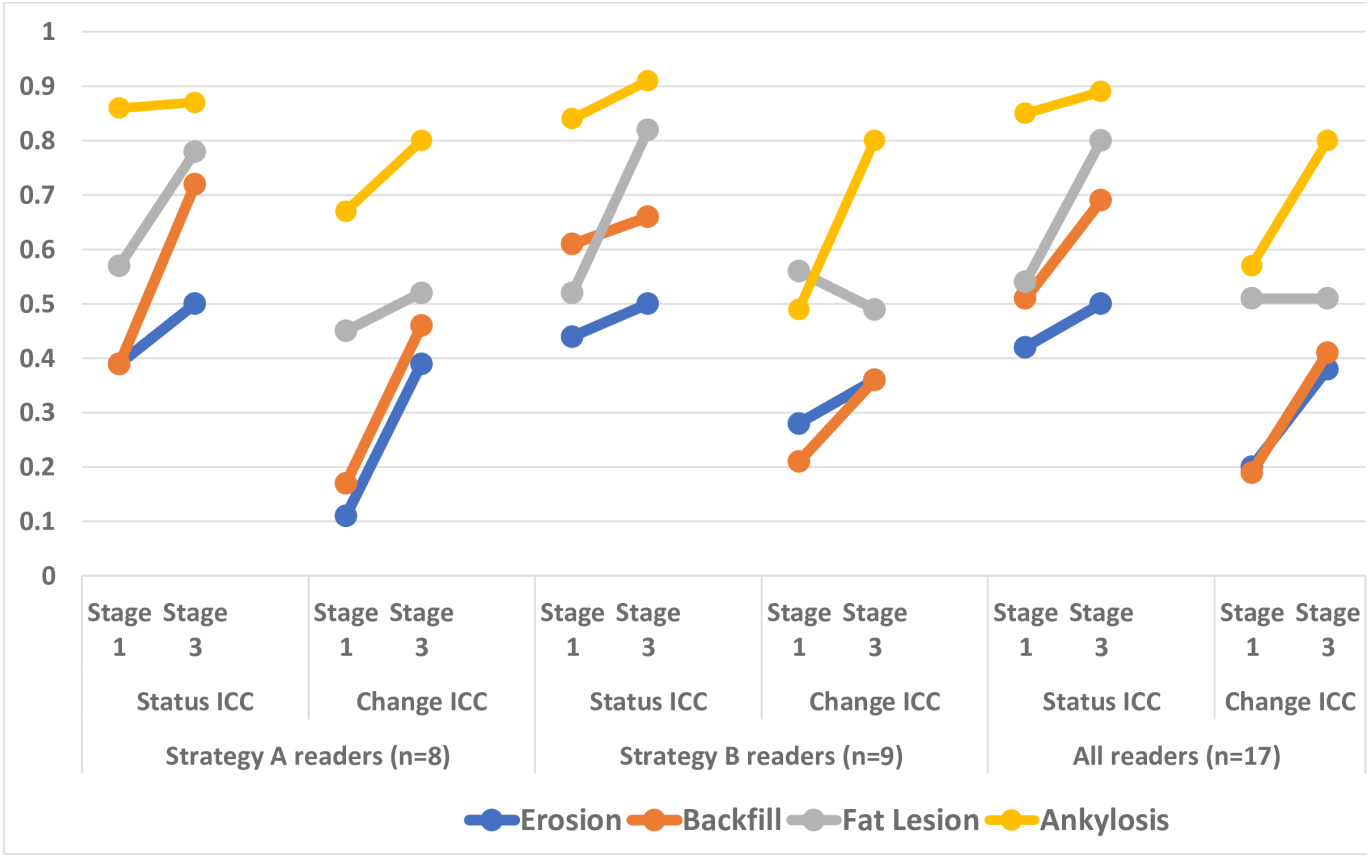

Figure 5 Reliability (mean intraclass correlation coefficient) between EuroSpA readers and Spondyloarthritis Research Consortium of Canada (SPARCC) developer radiologist for baseline and change in SPARCC structural scores for cases assessed after a review of the SPARCC manuscripts (stage 1) and after the completion of all calibration activities (stage 3). ICC, intraclass correlation coefficient.

that instruments should be feasible, but a framework for assessing the feasibility of imaging instruments has yet to be created. We have adhered to these recommendations in developing PowerPoint presentations for inflammatory and structural lesions in the SIJ that outline details of the scoring methodology and provide numerous examples. However, further training and calibration should include scans in DICOM format and from timepoints during which change in lesions might be expected when exposed to currently available therapies. This led to the

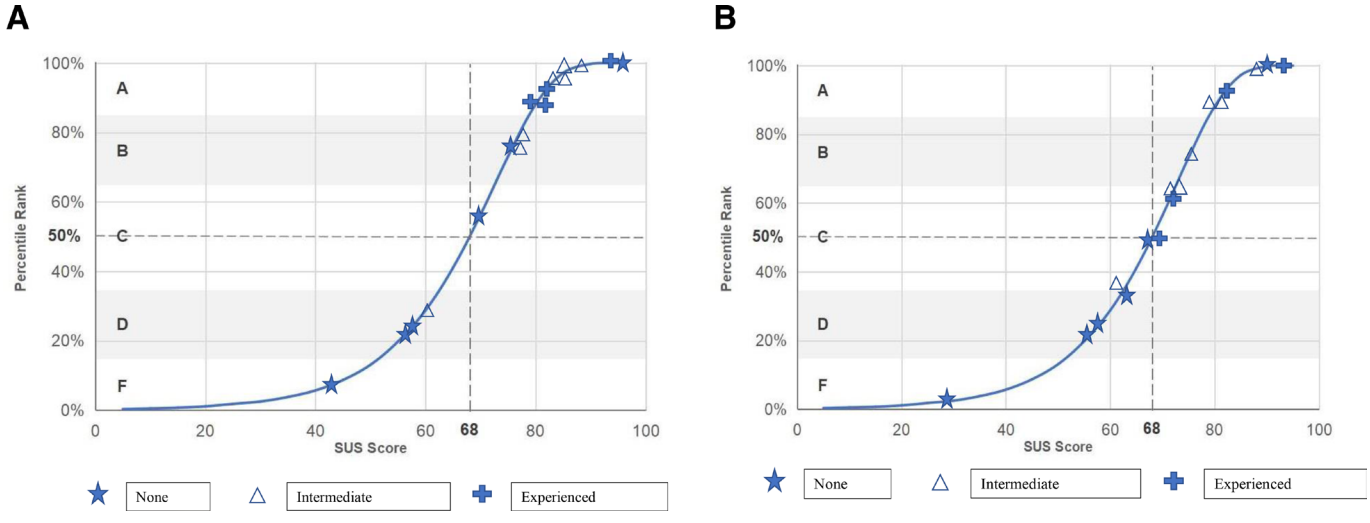

Figure 6 System Usability Scale scores and percentile ranking of SPARCC-SIJ$_{RETIC\text{-}INF}$, SPARCC-SIJ$_{RETIC\text{-}STR}$ modules by the level of prior expertise with Spondyloarthritis Research Consortium of Canada (SPARCC) methods. A score of 68 is generally regarded as the cut-off for an instrument likely to be widely applied. (A) SPARCC-SIJ$_{RETIC\text{-}INF}$ module. (B) SPARCC-SIJ$_{RETIC\text{-}STR}$ module.

additional development of the SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ modules. Such KT tools should be validated in terms of their feasibility and effectiveness in enhancing reader reliability.

Our data demonstrate that substantial training is necessary to score structural lesions with acceptable proficiency and that this can be enhanced with the KT tools that we have developed. This is unsurprising given the complex morphology of both erosions and backfill, the latter being defined on T1-weighted scans according to the presence of both complete loss of the dark appearance of the subchondral cortex at its expected location and an irregular band of dark signal reflecting sclerosis at the border of the original erosion.[32] Nevertheless, substantial enhancement of reliability was achieved for erosions and backfill after using the SPARCC-SIJ$_{RETIC\text{-}STR}$ module, irrespective of prior reader expertise with the SPARCC methods and with either strategy of calibration. The impact of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was more consistently observed for strategy B, particularly for fat lesions and ankylosis. Strategy A required readers to use the SPARCC-SIJ$_{RETIC\text{-}STR}$ module after scoring only one set of scans from 25 cases after a review of the manuscript describing the method. In comparison, strategy B required readers to score one set of scans from 25 cases after a review of the manuscript describing the method and a second set of 25 scans after a review of the PowerPoint presentation before the readers use the SPARCC-SIJ$_{RETIC\text{-}STR}$ module and then score the final set of scans from 25 cases. Consequently, strategy B entailed an additional training step before the SPARCC-SIJ$_{RETIC\text{-}STR}$ module was used, which could account for the more consistent impact on the reliability of this strategy. An alternative explanation may be provided by a review of the descriptive scores and reliability for SPARCC developers for stage 2 cases, which demonstrated a very small degree of change in backfill and substantially lower reliability for this lesion and to a lesser extent for ankylosis. While every attempt was made to ensure comparability in disease severity for the three different sets of scans, it appears likely that scans assessed at stage 2 were more complex. This could account for the less consistent impact of the SPARCC-SIJ$_{RETIC\text{-}STR}$ module when applied after readings of stage 1 cases compared with readings after stage 2 cases. Our finding that the reliability for structural lesion scores had improved from stage 1 to stage 3 after completion of all calibration activities and was comparable among readers irrespective of strategy or prior experience of the readers attests to the value of using a combination of PowerPoint and RETIC modules as KT tools.

An assessment of feasibility by a well-validated instrument, the SUS scale, supports the view that the SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$ modules have utility in enhancing learning and calibration, even for experienced readers. It is predictable that the lowest SUS scores would be observed for readers with no prior experience with the use of the SPARCC methods and that higher scores were observed with the SPARCC-SIJ$_{RETIC\text{-}INF}$ module as the assessment of BME is more straightforward than the assessment of structural lesions. SUS scores were also comparably high when readers were asked to rate the feasibility of the SPARCC methods indicating that the calibration modules reflected the ease of use of the SPARCC methods.

Study limitations include the small sample size of scans for each of the stages of assessment, which likely led to differences in the degree of severity of structural damage which may be a confounder in the interpretation of the impact of the calibration modules. Moreover, the evaluated cases had r-axSpA, often with concomitant lesions, as compared with early disease where lesions may have been more subtle. Reliability may vary with the extent and severity of the lesion, and we therefore cannot extrapolate our findings to nr-axSpA. Structural lesions were scored by viewing only the T1-weighted scans as compared with both the STIR and T1-weighted scans together, as is generally the case in routine practice. Simultaneous evaluation of different sequences enhances the interpretation of structural and inflammatory lesions and so it could be argued that the reliability data is overly conservative. However, the SPARCC scoring methods are primarily intended for use in clinical research of axSpA, especially clinical trials, where the simultaneous availability of STIR scans could unblind the reader to time sequence since substantial change in BME may be evident by the 12–16-week primary endpoint of placebo-controlled trials of axSpA. We also did not assess the long-term impact of the calibration modules on scoring proficiency, and it needs to be clarified how frequently readers should review the modules to maintain their scoring proficiency. It should be acknowledged that although the calibration modules enhanced scoring proficiency, there was still a substantial gap in the reliability attained by SPARCC developers, particularly for structural lesions, although some individual reader pairs did achieve reliability very comparable with SPARCC developers. This gap may be addressed by the future incorporation of additional MRI sequences that accentuate the signal contrast at the interface of the cartilage and bone and thereby enhance detection of erosion, such as three-dimensional gradient echo sequences with volumetric interpolated breath-hold examination.[33]

In conclusion, novel web-based calibration modules have been developed for the SPARCC MRI SIJ inflammation and structural scoring methods (SPARCC-SIJ$_{RETIC\text{-}INF}$ and SPARCC-SIJ$_{RETIC\text{-}STR}$) based on DICOM images, real-time iterative feedback and prespecified targets for attaining scoring proficiency. The modules, in combination with detailed PowerPoint ínstructions on pathologies and scoring methodology, enhanced scoring proficiency for the SPARCC MRI SIJ inflammation and structural methods in scoring exercises comprising 17 readers with varying expertise in these methods and 75 cases, each with pretreatment and post-treatment scans. The greatest enhancement of reader reliability was evident after using the SPARCC-SIJ RETIC-STR module, especially for

inexperienced readers, and was consistently evident for scoring erosion and backfill, even in experienced readers. The feasibility of both modules was evident by approximation of reading time per case with SPARCC developers after completion of calibration and by high SUS scores greater than the 50th percentile of normative data by the majority of readers. We therefore propose these modules for the routine calibration of readers prior to the use of these methods for clinical research and trials including MRI evaluation of the SIJ in patients with axSpA.

Author affiliations
[1]Medicine, University of Alberta, Edmonton, Alberta, Canada
[2]CARE ARTHRITIS Limited, Edmonton, Alberta, Canada
[3]Copenhagen Center for Arthritis Research (COPECARE), Center for Rheumatology and Spine Diseases, Rigshospitalet, Glostrup, Denmark
[4]Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
[5]Department of Rheumatology, University Hospital of Zurich, Zurich, Switzerland
[6]Copenhagen Center for Arthritis Research, Rigshospitalet, Copenhagen, Denmark
[7]Rheumatology, Geneva University Hospitals, Geneva, Switzerland
[8]First Faculty of Medicine, Rheumatology, Charles University, Prague, Czech Republic
[9]Rheumatology, University of Alberta Faculty of Medicine and Dentistry, Edmonton, Alberta, Canada
[10]Ghent University Hospital, Ghent, Oost-Vlaanderen, Belgium
[11]Leiden University, Leiden, Netherlands
[12]Center for Rheumatology and Spine Diseases, Rigshospitalet, Copenhagen, Denmark
[13]Clinical Immunology and Rheumatology, Christian Medical College and Hospital Vellore, Vellore, India
[14]Radiology, UKC Ljubljana, Ljubljana, Slovenia
[15]Radiology, Geneva University Hospitals, Geneva, Switzerland
[16]Rheumatology, Inselspital Universitatsspital Bern, Bern, Switzerland
[17]Diagnostic Imaging, Sheba Medical Center, Tel Hashomer, Israel
[18]Tel Aviv Sourasky Medical Center, Tel Aviv, Israel
[19]CARE Arthritis, Edmonton, Alberta, Canada
[20]Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Alberta, Canada

Twitter Walter Maksymowych @walter maks and Raphael Micheroli @ramicheroli

Contributors Substantial contributions to study conception and design: WM, AEFH, MØ, JP, RGWL. Substantial contributions to analysis and interpretation of the data: WM, AEFH, MØ, SW, JP, RGWL. Drafting the article or revising it critically for important intellectual content: WM, AEFH, MØ, RM, SJP, AC, NV, MSN, KB, SW, MdH, AJM, KP, MG, ZS, MW, KG, BM, IE, JP, RGWL. Final approval of the version of the article to be published: WM, AEFH, MØ, RM, SJP, AC, NV, MJN, KB, SW, MdH, AJM, KP, MG, ZS, MW, KG, BM, IE, JP, RGWL. WM is responsible for the overall content as guarantor. The guarantor accepts full responsibility for the finished work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests WM has received honoraria/consulting fees from AbbVie, BMS, Boehringer Ingelheim, Celgene, Eli Lilly, Galapagos, Janssen, Novartis, Pfizer and UCB Pharma; research grants from AbbVie, Pfizer and UCB Pharma; and educational grants from AbbVie, Janssen, Novartis and Pfizer. WM is the Chief Medical Officer for CARE ARTHRITIS. MØ has received research grants from AbbVie, BMS, Merck, Novartis and UCB and speaker and/or consultancy fees from AbbVie, BMS, Boehringer Ingelheim, Celgene, Eli Lilly, Galapagos, Gilead, Hospira, Janssen, MEDAC, Merck, Novartis, Novo, Orion, Pfizer, Regeneron, Roche, Sandoz, Sanofi and UCB. RM received honoraria for lectures or presentations from AbbVie, Eli Lilly, Janssen, Gilead and Pfizer. BM received travel expenditures, honoraria for lectures or presentations from AbbVie, Janssen, Novartis and Pfizer. MJN has received honoraria for travel expenditures, lectures or presentations from AbbVie, Eli Lilly, Janssen, Novartis, Pfizer and UCB. MdH received honoraria for presentations from UCB. RM received honoraria for presentations from UCB.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The e-tools are available free of charge for academic and not-for-profit entities. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The SPARCC MRI sacroiliac joint modules are accessible at: www.carearthritis.com/service/mri-scoring-modules/

ORCID iDs
Walter Maksymowych http://orcid.org/0000-0002-1291-1755
Anna Enevold Fløistrup E F Hadsbjerg http://orcid.org/0000-0001-8196-4327
Raphael Micheroli http://orcid.org/0000-0002-8918-7304
Susanne Juhl Pedersen http://orcid.org/0000-0002-6500-9263
Adrian Ciurea http://orcid.org/0000-0002-7870-7132
Michael S Nissen http://orcid.org/0000-0002-6326-1764
Manouk de Hooge http://orcid.org/0000-0002-0652-9808
Ashish J Mathew http://orcid.org/0000-0002-2061-2042
Marie Wetterslev http://orcid.org/0000-0002-2095-9441
Iris Eshed http://orcid.org/0000-0002-4655-9606

REFERENCES

1 Maksymowych WP. The role of imaging in the diagnosis and management of axial spondyloarthritis. *Nat Rev Rheumatol* 2019;15:657–72.
2 Maksymowych WP, Inman RD, Salonen D, *et al*. Spondyloarthritis research consortium of Canada magnetic resonance imaging index for assessment of sacroiliac joint inflammation in ankylosing spondylitis. *Arthritis Rheum* 2005;53:703–9.
3 Maksymowych WP, Inman RD, Salonen D, *et al*. Spondyloarthritis research consortium of canada magnetic resonance imaging index for assessment of spinal inflammation in ankylosing spondylitis. *Arthritis Rheum* 2005;53:502–9. 10.1002/art.21337 Available: http://doi.wiley.com/10.1002/art.v53:4
4 Landewé RBM, Hermann K-GA, van der Heijde DMFM, *et al*. Scoring sacroiliac joints by magnetic resonance imaging. A multiple-reader reliability experiment. *J Rheumatol* 2005;32:2050–5.
5 Lambert RGW, Rahman DSP, Inman RD, *et al*. Adalimumab significantly reduces both spinal and sacroiliac joint inflammation in patients with ankylosing spondylitis. *Arthritis Rheum* 2007;56:4005–14.
6 Maksymowych WP, Salonen D, Inman RD, *et al*. Low-dose infliximab (3 mg/kg) significantly reduces spinal inflammation on magnetic resonance imaging in patients with ankylosing spondylitis: a randomized placebo-controlled study. *J Rheumatol* 2010;37:1728–34.
7 Sieper J, van der Heijde D, Dougados M, *et al*. Efficacy and safety of adalimumab in patients with non-radiographic axial spondyloarthritis: results of a randomised placebo-controlled trial (ABILITY-1). *Ann Rheum Dis* 2013;72:815–22.
8 Dougados M, van der Heijde D, Sieper J, *et al*. Symptomatic efficacy of etanercept and its effects on objective signs of inflammation in early nonradiographic axial spondyloarthritis: a multicenter, randomized, double-blind, placebo-controlled trial. *Arthritis Rheumatol* 2014;66:2091–102.
9 Sieper J, van der Heijde D, Dougados M, *et al*. A randomized, double-blind, placebo-controlled, sixteen-week study of subcutaneous golimumab in patients with active nonradiographic axial spondyloarthritis. *Arthritis Rheumatol* 2015;67:2702–12. 10.1002/art.39257 Available: https://acrjournals.onlinelibrary.wiley.com/toc/23265205/67/10
10 Maksymowych WP, Dougados M, van der Heijde D, *et al*. Clinical and MRI responses to etanercept in early non-radiographic axial

spondyloarthritis: 48-week results from the EMBARK study. *Ann Rheum Dis* 2016;75:1328–35.
11 Braun J, Baraliakos X, Hermann K-G, *et al*. Effect of certolizumab pegol over 96 weeks of treatment on inflammation of the spine and sacroiliac joints, as measured by MRI, and the association between clinical and MRI outcomes in patients with axial spondyloarthritis. *RMD Open* 2017;3:e000430.
12 van der Heijde D, Deodhar A, Wei JC, *et al*. Tofacitinib in patients with ankylosing spondylitis: a phase II, 16-week, randomised, placebo-controlled, dose-ranging study. *Ann Rheum Dis* 2017;76:1340–7.
13 van der Heijde D, Baraliakos X, Hermann K-GA, *et al*. Limited radiographic progression and sustained reductions in MRI inflammation in patients with axial spondyloarthritis: 4-year imaging outcomes from the RAPID-axSpA phase III randomised trial. *Ann Rheum Dis* 2018;77:699–705.
14 van der Heijde D, Sieper J, Maksymowych WP, *et al*. Clinical and MRI remission in patients with nonradiographic axial spondyloarthritis who received long-term open-label adalimumab treatment: 3-year results of the ABILITY-1 trial. *Arthritis Res Ther* 2018;20:61.
15 van der Heijde D, Baraliakos X, Gensler LS, *et al*. Efficacy and safety of filgotinib, a selective Janus kinase 1 inhibitor, in patients with active ankylosing spondylitis (TORTUGA): results from a randomised, placebo-controlled, phase 2 trial. *Lancet* 2018;392:2378–87.
16 Maksymowych WP, van der Heijde D, Baraliakos X, *et al*. Tofacitinib is associated with attainment of the minimally important reduction in axial magnetic resonance imaging inflammation in ankylosing spondylitis patients. *Rheumatology (Oxford)* 2018;57:1390–9.
17 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.
18 van der Heijde D, Song I-H, Pangan AL, *et al*. Efficacy and safety of upadacitinib in patients with active ankylosing spondylitis (SELECT-AXIS 1): a multicentre, randomised, double-blind, placebo-controlled, phase 2/3 trial. *Lancet* 2019;394:2108–17.
19 Deodhar A, Gensler LS, Kay J, *et al*. A fifty-two-week, randomized, placebo-controlled trial of certolizumab pegol in nonradiographic axial spondyloarthritis. *Arthritis Rheumatol* 2019;71:1101–11.
20 Deodhar A, van der Heijde D, Gensler LS, *et al*. Ixekizumab treatment in patients with non-radiographic axial spondyloarthritis: results of a 52-week, randomized. *Lancet (in Press)*
21 Weiss PF, Maksymowych WP, Lambert RG, *et al*. Feasibility and reliability of the spondyloarthritis research consortium of Canada sacroiliac joint inflammation score in children. *Arthritis Res Ther* 2018;20:56.
22 Lukas C, Braun J, van der Heijde D, *et al*. Scoring inflammatory activity of the spine by magnetic resonance imaging in ankylosing spondylitis: a multireader experiment. *J Rheumatol* 2007;34:862–70.
23 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.
24 Maksymowych WP, Wichuk S, Chiowchanwisawakit P, *et al*. Development and preliminary validation of the spondyloarthritis research consortium of Canada magnetic resonance imaging sacroiliac joint structural score. *J Rheumatol* 2015;42:79–86.
25 Maksymowych WP, Wichuk S, Dougados M, *et al*. Modification of structural lesions on MRI of the sacroiliac joints by etanercept in the EMBARK trial: a 12-week randomised placebo-controlled trial in patients with non-radiographic axial spondyloarthritis. *Ann Rheum Dis* 2018;77:78–84.
26 Maksymowych WP, Østergaard M, Landewé R, *et al*. Impact of filgotinib on sacroiliac joint magnetic resonance imaging structural lesions at 12 weeks in patients with active ankylosing spondylitis (TORTUGA trial). *Rheumatology (Oxford)* 2022;61:2063–71.
27 Maksymowych WP, Baraliakos X, Lambert RG, *et al*. Effects of ixekizumab treatment on structural changes in the sacroiliac joint: MRI assessments at 16 weeks in patients with non-radiographic axial spondyloarthritis. *The Lancet Rheumatology* 2022;4:e626–34.
28 Brooke J. “SUS: a "quick and dirty" usability scale”. In: Jordan PW, Thomas B, Weerdmeester BA, et al., eds. *Usability Evaluation in Industry*. London: Taylor and Francis, 1996.
29 Sauro J. *A Practical Guide to the System Usability Scale: Background, Benchmarks*. Denver, CO: Measuring Usability LLC, 2011.
30 Lewis JR, Sauro J. Item benchmarks for the system usability scale. *J Usability Stud* 2018;13:158–67.
31 D’Agostino MA, Beaton DE, Maxwell LJ, *et al*. Improving domain definition and outcome instrument selection: lessons learned for OMERACT from imaging. *Semin Arthritis Rheum* 2021;51:1125–33.
32 Maksymowych WP, Lambert RG, Østergaard M, *et al*. MRI lesions in the sacroiliac joints of patients with spondyloarthritis: an update of definitions and validation by the ASAS MRI working group. *Ann Rheum Dis* 2019;78:1550–8.
33 Diekhoff T, Greese J, Sieper J, *et al*. Improved detection of erosions in the sacroiliac joints on MRI with volumetric interpolated breath-hold examination (VIBE): results from the SIMACT study. *Ann Rheum Dis* 2018;77:1585–9.