Scoring Sacroiliac Joints by Magnetic Resonance Imaging. A Multiple-Reader Reliability Experiment

ROBERT B.M. LANDEWÉ, KAY-GEERT A. HERMANN, DÉSIRÉE M.F.M. VAN DER HEIJDE, XENOPHON BARALIAKOS, ANNE-GRETHE JURIK, ROBERT G. LAMBERT, MIKKEL ØSTERGAARD, MARTIN RUDWALEIT, DAVID C. SALONEN, JÜRGEN BRAUN, and the ASAS/OMERACT MRI in AS Working Group

ABSTRACT. Magnetic resonance imaging (MRI) of the sacroiliac (SI) joints and the spine is increasingly important in the assessment of inflammatory activity and structural damage in clinical trials with patients with ankylosing spondylitis (AS). We investigated inter-reader reliability and sensitivity to change of several scoring systems to assess disease activity and change in disease activity in patients with AS. Twenty sets of consecutive MRI, derived from a randomized clinical trial comparing an active drug with placebo and selected on the basis of the presence of activity at baseline, were presented electronically to 7 experienced readers from different countries (Europe, Canada). Readers scored the MRI by 3 different methods including: a global score (grading activity per SI joint); a more comprehensive global score (grading activity per SI joint per quadrant); and a detailed scoring system [Spondyloarthritis Research Consortium of Canada (SPARCC) scoring system], which scores 6 images, divided into quadrants, with additional scores for "depth" and "intensity." A fourth and a fifth scoring system were constructed afterwards. The fourth method included the SPARCC score minus the additional scores for "depth" and "intensity," and the fifth method included the SPARCC slice with the maximum score. Inter-reader reliability was investigated by calculating intraclass correlation coefficients (ICC) for all readers together and for all possible reader pairs. Sensitivity to change was investigated by calculating standardized response means (SRM) on change scores that were made positive. Overall inter-reader ICC per method were between 0.47 and 0.58 for scoring status, and between 0.40 and 0.53 for scoring change. ICC per possible reader pairs showed much more fluctuation per method, with lowest observed values close to zero (no agreement) and highest observed values over 0.80 (excellent agreement). In general, agreement of status scores was somewhat better than agreement of change scores, and agreement of the comprehensive SPARCC scoring system was somewhat better than agreement of the more condensed systems. Sensitivity to change differed per reader, but in general was somewhat better for the comprehensive SPARCC system. This experiment under "real life," far from optimal conditions demonstrates the feasibility of scoring exercises for method comparison, provides evidence for the reliability and sensitivity to change of scoring systems to be used in assessing activity of SI joints in clinical trials, and sets the conditions for further validation research in this field. (J Rheumatol 2005;32:2050–5)

Key Indexing Terms:
ANKYLOSING SPONDYLITIS MAGNETIC RESONANCE IMAGING
SACROILIAC JOINT VALIDITY RELIABILITY

Introduction

Ankylosing spondylitis (AS), a chronic debilitating inflammatory rheumatic disease, affects the sacroiliac (SI) joints. Inflammation of the SI joints leads to the formation of erosions, sclerosis, bony bridging, and complete ankylosis. The processes that interfere with the bony delineation of the SI

From the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht, Maastricht, The Netherlands; Departments of Radiology and Rheumatology, Charité Medical School, Berlin, Germany; Rheumazentrum Ruhrgebiet, Herne, Germany; Department of Radiology, Aarhus University Hospital, Aarhus, Denmark; Department of Radiology, University of Alberta, Edmonton, Canada; University of Copenhagen Hvidovre Hospital, Hvidovre, Denmark; and Department of Rheumatology, University of Toronto, Toronto, Canada.

R. Landewé, MD, PhD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht; K-G. Hermann, MD, Department of Radiology, Charité Medical School; D.M.F.M. van der Heijde, MD, PhD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht; X. Baraliakos, Rheumazentrum Ruhrgebiet; A-G. Jurik, Department of Radiology, Aarhus University Hospital; R.G. Lambert, Department of Radiology, University of Alberta; M. Østergaard, MD, PhD, DMSc, University of Copenhagen, Hvidovre Hospital; M. Rudwaleit, MD, Department of Rheumatology Charité — Campus Benjamin Franklin; D.C. Salonen, Department of Rheumatology, University of Toronto; J Braun, Rheumazentrum Ruhrgebiet.

The ASAS/OMERACT MRI in AS Working Group: Jürgen Braun, Désirée van der Heijde (chairs), Xenofon Baraliakos, Matthias Bollow, Paul Emery, Kay-Geert Hermann, Robert Inman, Anne-Grethe Jurik, Mart van de Laar, Robert G. Lambert, Robert Landewé, Walter Maksymowych, Helena Marzo-Ortega, Phil O'Connor, Mikkel Østergaard, Ans Oostveen, Martin Rudwaleit, David Salonen, Jochen Sieper, Millicent Stone, and Kurt de Vlam.

Address reprint requests to Dr. R. Landewé, Department of Internal Medicine/Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands. E-mail: rlan@sint.azm.nl

joints (erosions, bridging) can be detected on radiographs. Bony changes of the SI joints are sufficiently specific for inflammatory spondyloarthropathy that they have been included in classification criteria for AS used in research. But bony changes, often referred to as structural changes, occur relatively late in the course of the disease and do not measurably change over short periods of time. The late development of structural changes renders them unsuitable for use in early diagnosis and their slow evolution prevents their use as outcome variables in short term clinical trials.

Inflammation itself cannot be detected on radiographs. However, the development of magnetic resonance imaging (MRI) has allowed inflammation to be visualized in SI joints in acute stages, with implications for earlier diagnosis of AS. Since inflammation appears to be a dynamic process that can be easily identified with MRI, investigators have recommended using sequential MRI to measure inflammatory activity of the SI joints as an outcome measure in clinical trials.

One requirement for using scoring methods in clinical trials is that they pass the OMERACT filter[1]. Two important aspects of discrimination are interobserver reliability and sensitivity to change over time. The ASsessment in Ankylosing Spondylitis/OMERACT MRI (ASAS/OMERACT MRI) working group for MRI in AS[2] decided to take an inventory of all available and potential scoring methods for inflammation of SI joints, and to test comparative reliability of these scoring methods, as well as comparative sensitivity to change.

Methods

Process. A total of 6 potentially useful scoring methods were identified: the MR Imaging of Seronegative SpA (MISS) scoring system[3], an unpublished Leeds scoring system, the Aarhus scoring system[4], the Spondyloarthritis Research Consortium of Canada (SPARCC) scoring system[5], and 2 initiatives from Berlin, one published as a proposal[6] and one unpublished. Scoring systems differed with respect to the MRI sequence required to detect inflammation [short-tau inversion recovery (STIR), T1, T1/Gd, T2 with fat suppression], the unit of interest (SI joint divided into quadrants or halves), the number of slices that were scored, the slice orientation (coronal oblique, sagittal, semi-coronal/semi-axial), the measurement feature used to assess inflammatory lesions (global grading, extent, intensity), the site of the inflammatory lesion to be scored (subchondral bone, bone marrow, joint space), and as a consequence of all these facts together, the range of the scoring system.

The different scoring systems were discussed with respect to all these aspects, and for performance in terms of validity. There was broad consensus in the group that an extensive method should be compared with a global one, that different qualities, such as extent and depth of the inflammatory lesion, should be taken into account, and that the additional value (and/or additional error) of scoring different slices should be tested.

It was decided that the SPARCC scoring system should serve as a template to test inter-reader reliability in a first experiment, because it is the most comprehensive scoring system of all 6 systems that were judged: it includes additional features of potential importance such as extent and depth of inflammation, and uses most MRI slices to score. It was also decided that the SPARCC scoring system should be compared to a global graded scoring system, in order to test the influence of comprehensiveness on inter-reader variability, and that the influence of the additional qualities "extent" and "depth" on inter-reader variability should be tested.

The SPARCC Scoring System

This scoring method is based entirely on the assessment of increased signal on T2 with fat suppression or STIR sequences denoting bone marrow edema on oblique coronal slices of the SI joint. All such signal changes within the iliac bone and sacrum up to the sacral foramina are scored on 6 consecutive slices through the SI joint. These slices are selected based on the SPARCC protocol, which defines acquisition characteristics, and they encompass most of the synovial compartment. Sacral interforaminal bone marrow STIR signal forms the reference for determination of increased signal in the SI joints. Each SI joint is divided into 4 quadrants, and the presence of increased STIR signal in each of these 4 quadrants is recorded in each of the 6 slices, giving a maximum score of 48. The presence of a lesion exhibiting either intense signal (comparable to signal from adjacent blood vessels) or depth $\geq$ 1 cm anywhere within each SI joint of the 6 slices is given an additional score (i.e., a maximum additional score of 4 per slice), bringing the total score to 72.

Experiment to Test Inter-Reader Variability and Sensitivity to Change

Twenty sets of MRI were sent out electronically to 8 experienced readers, all members of the ASAS/OMERACT MRI in AS working group. Seven readers returned complete databases and their results were included. The results from the eighth reader were omitted because of too many missing data.

Three readers were rheumatologists and 4 were radiologists. Sets of MRI were selected by 2 of us who did not take part in reading (RL, DH), from a randomized clinical trial comparing an active drug with placebo. Selection was performed in such a way that patients with and without inflammatory activity of the SI joints were included in a balanced way. One MRI set consisted of baseline images (T1, STIR, T1/gadolinium) as well as images performed after treatment (placebo or active).

Readers followed instructions in the form of written guidelines describing the different scoring systems. Readers were not otherwise trained specifically for this exercise.

Readers had to: (1) give a global grading [from 0 (no inflammation) to 3 (most extensive inflammation)] of both SI joints separately (Method 1; range of total score 0 to 6); (2) give a global graded impression of both SI joints per quadrant (Method 2; range of total score from 0 to 24); and (3) perform the SPARCC score (see above) (Method 3; range of total score from 0 to 72). Readers then assigned by number which slice they started scoring, and which slice showed the most extensive inflammation according to their judgment. Readers were blinded to trial treatment and time sequence of the images. Scores were filled in electronically (predesigned Excel sheet), and forwarded to one center for further analysis.

Analysis

Data were aggregated and analyzed by one of us (RL). Before further inference, 2 additional scoring methods were derived from Method 3: the SPARCC scoring system, but with a sum score calculated without additional scores for "depth" and "extent" (Method 4; range of total score from 0 to 48); and the slice with maximum score (Method 5; range of total score from 0 to 12).

Inter-reader variability was determined per scoring method by intraclass correlation coefficient (ICC, absolute agreement definition) for all readers together, for every reader pair, and for radiologists and rheumatologists separately, for both status scores (baseline values) and change scores.

Sensitivity to change was assessed by calculating standardized response means per reader per method on "absolute" change scores (both positive and negative changes were treated as change without taking the direction into account, since the treatment code was not available and patients could have worsened or improved).

Results

Seven of 8 readers provided complete scoring sheets (20 patients, 2 timepoints). Table 1 shows the grand means of the scorings per method (all patients, all readers) as well as the range of scores that were observed in the experiment. It is obvious that the grand means are all in the lower half of the entire scoring range, and that the global scores with limited range (Method 1, Method 2, Method 5) on average use a proportion of the range greater than the comprehensive scores (Method 3, Method 4).

Inter-reader variability per method is provided in Table 2. Overall ICC incorporate and aggregate all sources of variability among different readers, and were around 0.50 for all 5 methods, for both status scores and change scores, with only small differences between methods.

ICC calculated per reader pair for both status score and change score are summarized in Table 3 according to method and their range (lowest observed reader pair ICC; highest observed reader pair ICC). In contrast to what the global ICC in Table 2 suggest, there is a lot of variability in agreement among different reader pairs, for both status scores and change scores. The range between lowest and highest inter-reader ICC is somewhat greater for change scores compared to status scores. For both status scores and change scores, the lowest observed inter-reader ICC is found more than once for the same reader pair (R3;R5 and R5;R6 for status scores, and R6;R8 for change scores). Reader pairs R1;R7 (2 times) and R2;R6 (3 times) demonstrated the highest observed inter-reader ICC, for status score and change score, respectively.

SRM according to method and reader are summarized in Table 4. The matrix shows that the SRM show important variation, both per reader and per method. If the median SRM of 7 readers is considered representative for the performance of a method, the SRM of Method 3 (SPARCC) is somewhat higher than the SRM of the other methods. If the median SRM of 5 different methods is considered representative for the performance of a reader, it is obvious that some readers do not see much change, whereas others do. It should be noted that there is large variation in SRM, with reader 1 obtaining the highest median value, and reader 3 obtaining the lowest median value.

Discussion

The results of this study give rise to different conclusions. (1) We have shown that it is feasible to perform inter-reader reliability experiments with MRI with a large number of readers located around the world, by making use of electronic dissemination. (2) Although, overall, ICC show a

Table 1. Descriptive data.

	Grand Mean (observed range)	
	First Time Point	Second Time Point
Method 1: Global score 0 to 6	2.1 (0 to 6)	1.6 (0 to 6)
Method 2: Global quadrants 0 to 24	5.1 (0 to 20)	3.6 (0 to 21)
Method 3: SPARCC* 0 to 72	17.4 (0 to 61)	11.7 (0 to 67)
Method 4: SPARCC "minus"** 0 to 48	12.9 (0 to 38)	9.2 (0 to 43)
Method 5: SPARCC "max" # 0 to 12	4.4 (0 to 11)	3.3 (0 to 12)

* SPARCC: Spondyloarthritis Research Consortium of Canada; ** SPARCC minus the additional scores for "depth" and "intensity"; # SPARCC slice with highest score.

Table 2. Inter-reader reliability: Overall (all 7 readers).

	Intraclass Correlation Coefficient (95% CI)	
	On Status Scores (calculated on 1st time point)	On Change Scores
Method 1: Global score	0.49 (0.29–0.70)	0.40 (0.22–0.63)
Method 2: Global quadrants	0.48 (0.27–0.70)	0.53 (0.35–0.73)
Method 3: SPARCC*	0.55 (0.34–0.75)	0.52 (0.34–0.72)
Method 4: SPARCC "minus"**	0.47 (0.27–0.69)	0.52 (0.33–0.72)
Method 5: SPARCC "max" #	0.58 (0.37–0.77)	0.48 (0.29–0.69)

* SPARCC: Spondyloarthritis Research Consortium of Canada; ** SPARCC minus the additional scores for "depth" and "intensity"; # SPARCC slice with highest score.

Table 3. Inter-reader reliability: Reader pairs.

	Lowest Observed Inter-reader ICC (reader pair)	Highest Observed Inter-reader ICC (reader pair)	No. of Reader Pairs (%) with ICC ≥ 0.60	No. of Reader Pairs (%) with ICC ≥ 0.80
A. Status scores				
Method 1: Global score	0.19 (R6;R7)	0.82 (R3;R5)	7 (33)	1 (5)
Method 2: Global quadrants	0.39 (R3;R6)	0.82 (R2;R3)	12 (57)	1 (5)
Method 3: SPARCC	0.30 (R3;R6)	0.85 (R1;R8)	16 (76)	5 (24)
Method 4: SPARCC "minus"	0.13 (R3;R6)	0.86 (R1;R8)	14 (67)	3 (14)
Method 5: SPARCC "max"	0.43 (R6;R7)	0.84 (R2;R7)	15 (71)	7 (33)
B. Change scores				
Method 1: Global score	–0.12 (R5;R7)	0.57 (R2;R3)	0 (0)	0 (0)
Method 2: Global quadrants	0.11 (R5;R7)	0.86 (R1;R4)	2 (10)	1 (5)
Method 3: SPARCC	0.27 (R5;R7)	0.89 (R2;R76)	4 (19)	1 (5)
Method 4: SPARCC "minus"	0.20 (R1;R3)	0.85 (R2;R6)	2 (10)	1 (5)
Method 5: SPARCC "max"	0.12 (R1;R6)	0.77 (R2;R6)	2 (10)	0 (0)

Table 4. Sensitivity to change per reader and per scoring method.

	R1*	R2	R3	R4	R5	R6	R7	Median
Method 1: Global score	0.72**	0.72	0.34	0.50	0.41	0.53	0.49	0.50
Method 2: Global quadrants	0.56	0.68	0.47	0.67	0.39	0.64	0.46	0.56
Method 3: SPARCC	0.96	0.81	0.26	0.86	0.36	0.70	0.69	0.70
Method 4: SPARCC "minus"	0.88	0.72	0.17	0.77	0.20	0.61	0.62	0.62
Method 5: SPARCC "max"	0.89	0.56	0.24	0.82	0.17	0.51	0.59	0.56
Median per reader	0.88	0.72	0.26	0.77	0.36	0.61	0.59	

* Reader 1; ** Values reflect standardized response means calculated over 20 patients.

moderately good level of agreement among readers, inter-reader variability in scoring SI joint activity among different not specifically trained readers appears to be substantial regardless of the specific scoring method used. (3) Inter-reader variability across different methods (global as well as comprehensive) does not differ so importantly that a preference can be made on the basis of this aspect of reliability. (4) Last, notwithstanding significant variability among readers, sensitivity to change of the comprehensive SPARCC scoring system seems to be better than that of the "condensed" SPARCC systems, or of the more global scoring systems.

The feasibility of this experiment is important to mention. We decided to perform the scoring experiment, designed scoring rules and scoring sheet, and disseminated the sets of MRI to the readers who expressed their interest, and they in turn had to score the entire set within 2 weeks. Only one of 8 selected readers was not successful in completing this task within the narrow timeframe, and the scorings of this reader were excluded. Undoubtedly, such an experiment is only possible if MRI and scorings can be disseminated electronically, making use of specific software. Such an approach may also have disadvantages, of which loss of image quality (spatial resolution) is potentially the most important. Such loss of quality is at the cost of sensi-

tivity to change (lower SRM), and may be a source of additional noise (lower ICC). There are many more sources of variability that may jeopardize reliability in this experiment, such as the difference in training level (heterogeneity in reader performance), the selection of the images, and others. Therefore, the ICC and SRM that this experiment provided may be considered as conservative, reflecting "real life" rather than "optimal circumstances."

As mentioned, the overall ICC were moderately high, with values of about 0.50. The translation of these ICC values into an understandable concept is that 50% of all variation generated in this experiment by all potential sources was due to variation among patients. Because all readers saw the same patients, this should be the only true source of variation. The important other source of variation in this experiment was the reader. There are a number of reasons for the relatively high level of inter-reader variability: Readers were not trained with regard to the methods used here; there was no definition of abnormalities or lesions; sets of MRI were not prepared for a specific scoring system; quality of imaging was not optimal; and readers included both radiologists and rheumatologists; and others. In view of all these limitations, overall ICC of about 0.50 should be considered a good starting point for further research.

Somewhat unexpectedly, ICC across different methods did not differ importantly. A preference on the basis of inter-reader reliability could not be made. This means that the "quick" global systems, with ranges from 0 to 3 or 0 to 6, were about as reliable in terms of inter-reader variation as the comprehensive and time-consuming SPARCC system, with ranges from 0 to 72. However, the comprehensiveness of the SPARCC method may indicate an increased requirement for training in the scoring of this method.

Another important finding was that overall ICC obscured a high level of variability in the performances of different reader pairs. It was obvious that on the basis of these results "good" and "bad" reader pairs could be formed that consistently (i.e., with different methods) performed worse or better than the remainder. This finding stresses the importance of not randomly selecting readers for scoring clinical trial MRI, because statistical power in a clinical trial is — among many other factors — dependent on inter-reader reliability of change scores, which preferably should exceed an ICC of 0.80. In this study, with minimal training, and under the far from optimal conditions presented here, almost no reader pair met this criterion. However, looking at the percentage of possible reader pairs that meet a certain threshold, with regard to scoring both status and change, a certain preference for the comprehensive SPARCC system can be deducted, because a greater proportion of possible reader pairs meets the threshold of 0.60. The probable explanation is that sum scores such as the SPARCC efface a lot of inter-reader variability at the level of small units.

Sensitivity to change was only partially investigated. The absolute values of the SRM should be interpreted with extreme caution, because we did not know whether patients were receiving active treatment or placebo. Moreover, the set of MRI presented to the readers was only a selection, not an entire treatment group. But because the same set of patients was scored with different methods, the SRM can be used to make inter-method comparisons. When this was done, our results indicate a slight preference for the comprehensive SPARCC method. Five of the 7 readers reached highest SRM with the SPARCC system. Generally, comprehensive scoring systems with many levels have higher sensitivity to change, simply because they have a higher number of units to which change can be attributed by the reader. It also appeared from this experiment that there are readers that are very sensitive to change, in contrast to readers that hardly show change.

A concern with the design of this study is that scores for the various methods are collected in the same reading session. This means that contamination between scoring methods cannot be ruled out. In order to minimize contamination, it was intended that readers apply the global scoring method first, followed by a more detailed scoring by each additional method. Unaware of this issue in advance, one reader admitted afterwards that he changed global scores based on the additional scoring methods; however, excluding data from this reader did not influence the results or the conclusions.

To summarize, in the context of the international ASAS/OMERACT MRI in AS working group, we successfully performed an MRI reading exercise of the SI joints. Five methods were investigated with respect to inter-reader reliability and sensitivity to change. Inter-reader reliability was moderate on average, better for scoring status compared to change, and included significant variability in performance across different possible reader pairs. The chosen methods performed similarly, although more possible reader pairs reached acceptable ICC with the comprehensive SPARCC scoring system. Sensitivity to change was slightly better with the SPARCC system compared to the global scoring systems or the condensed ones.

It will be important now to reanalyze reliability in the optimal context, i.e., every developer of a scoring method tests reliability and sensitivity to change of their own method, so that it will become obvious whether experience and specific training will improve performance.

REFERENCES

1. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology [editorial]. J Rheumatol 1998;25:198-9.
2. van der Heijde D, Landewé R, Hermann KG, et al. Application of the OMERACT filter to scoring methods for MRI of the SI joints and the spine. Recommendations for a research agenda at OMERACT 7. J Rheumatol 2005;32:2048-9.
3. Marzo-Ortega H, Braun J, Maksymowych W, et al. Interreader agreement in the assessment of magnetic resonance imaging of the sacroiliac joints in spondyloarthropathy — the 1st MISS study

[abstract]. Arthritis Rheum 2002;46 Suppl:S428.
4. Puhakka KB, Jurik AG, Egund N, et al. Imaging of sacroiliitis in early seronegative spondylarthropathy. Assessment of abnormalities by MR in comparison with radiography and CT. Acta Radiol 2003;44:218-29.
5. Maksymowych WP, Dhillon SS, Inman RD, et al. The Spondyloarthritis Research Consortium of Canada (SPARCC) Magnetic Resonance Imaging (MRI) Index: A new scoring system for the evaluation of sacroiliac joint inflammation in spondyloarthritis [abstract]. Ann Rheum Dis 2004;63 Suppl:76.
6. Hermann KG, Braun J, Fischer T, Reisshauer BH, Bollow M. Magnetic resonance tomography of sacroiliitis: anatomy, histological pathology, MR-morphology, and grading [German]. Radiologe 2004;44:217-28.