Scoring Inflammatory Activity of the Spine by Magnetic Resonance Imaging in Ankylosing Spondylitis: A Multireader Experiment

CÉDRIC LUKAS, JÜRGEN BRAUN, DÉSIRÉE van der HEIJDE, KAY-GEERT A. HERMANN, MARTIN RUDWALEIT, MIKKEL ØSTERGAARD, ANS OOSTVEEN, PHIL O'CONNOR, WALTER P. MAKSYMOWYCH, ROBERT G.W. LAMBERT, ANNE GRETHE JURIK, XENOFON BARALIAKOS, ROBERT LANDEWÉ, for the ASAS/OMERACT MRI in AS Working Group

ABSTRACT. ***Objective.*** Magnetic resonance imaging (MRI) of the spine is increasingly important in the assessment of inflammatory activity in clinical trials with patients with ankylosing spondylitis (AS). We investigated feasibility, inter-reader reliability, sensitivity to change, and discriminatory ability of 3 different scoring methods for MRI activity and change in activity of the spine in patients with AS.

Methods. Thirty sets of spinal MRI at baseline and after 24 weeks of followup, derived from a randomized clinical trial comparing a tumor necrosis factor (TNF)-blocking drug (n = 20) with placebo (n = 10) and selected to cover a wide range of activity at baseline and change in activity, were presented electronically in a partial latin-square design to 9 experienced readers from different countries (Europe, Canada). Readers scored each set of MRI 3 times, using 3 different methods including the Ankylosing Spondylitis spine Magnetic Resonance Imaging-activity [ASspiMRI-a, grading activity (0-6) per vertebral unit in 23 units]; the Berlin modification of the ASspiMRI-a; and the Spondyloarthritis Research Consortium of Canada (SPARCC) scoring system, which scores the 6 vertebral units considered by the reader as the most abnormal, with additional scores for "depth" and "intensity." Both the order of the methods used by each reader and the timepoints (before/after treatment) were randomized. Feasibility of each scoring system was evaluated by measuring the mean time needed to score each set of MRI, and inter-reader reliability was evaluated by smallest detectable change (SDC) and by intraclass correlation coefficients (ICC) for all readers together and for all possible reader pairs separately. Sensitivity to change was investigated by calculating Guyatt's effect size on change scores. Discriminatory ability was assessed using Z-scores (Mann-Whitney test) comparing change in score between patients treated with TNF-blocking drug and placebo.

Results. The mean time to score one set of MRI was shortest for the Berlin method. SDC was lowest for the Berlin method and highest for SPARCC. Overall inter-reader ICC per method were between 0.49 and 0.77 for scoring activity status, and between 0.46 and 0.72 for scoring activity change. ICC for all possible reader pairs showed much more fluctuation per method, with lowest observed values of about 0.05 (very low agreement) and highest observed values over 0.90 (excellent agreement). In general, ICC for SPARCC were consistently higher than for other systems. Sensitivity to change differed per reader, and was more consistent with SPARCC than with the other methods, but was in general excellent for all 3 methods. Discrimination between groups (TNF-blocker vs placebo) assessed by Z-scores was good and comparable among methods.

Conclusion. This experiment demonstrates the feasibility of multiple-reader MRI scoring exercises for method comparison, provides evidence for the feasibility, reliability, sensitivity to change, and discriminatory capacity of all 3 tested scoring systems to be used in assessing spinal activity on MRI in patients with AS in clinical trials. On the basis of these results it is not possible to prioritize one of the 3 methods. (J Rheumatol 2007;34:862–70)

Key Indexing Terms:
MAGNETIC RESONANCE IMAGING
ANKYLOSING
SPONDYLITIS
VALIDATION STUDIES

From the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht; CAPHRI Research Institute, University Maastricht; Department of Rheumatology, Twenteborg Hospital, Almelo, the Netherlands; University of Copenhagen, Hvidovre Hospital, Copenhagen; Department of Radiology, Aarhus University Hospital, Denmark; Rheumazentrum Ruhrgebiet, Herne; Department of Rheumatology Charité - Campus Benjamin Franklin; Department of Radiology, Charité Medical School, Berlin, Germany; Department of Radiology, Leeds General Infirmary, UK; and Department of Medicine, University of Alberta, Edmonton, Alberta, Canada.

Dr. Maksymowych is a Senior Scholar of the Alberta Heritage Foundation for Medical Research.

C. Lukas, MD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht; J. Braun, MD, PhD, Rheumazentrum Ruhrgebiet; D. van der Heijde, MD, PhD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht and CAPHRI Research Institute, University Maastricht;

M. Østergaard, MD, PhD, DMSc, University of Copenhagen, Hvidovre Hospital; A. Oostveen, Department of Rheumatology, Twenteborg Hospital; M. Rudwaleit, MD, Department of Rheumatology Charité - Campus Benjamin Franklin;
P. O'Connor, MD, Radiologist, Department of Radiology, Leeds General Infirmary; W.P. Maksymowych, FRCPC, Professor of Medicine, Department of Medicine; R.G.W. Lambert, MB, FRCPC, Associate Professor of Radiology, University of Alberta; A.G. Jurik, MD, DMSc, Department of Radiology, Aarhus University Hospital; K-G.A. Hermann, MD, Department of Radiology, Charité Medical School; X. Baraliakos, Rheumazentrum Ruhrgebiet; R. Landewé, MD, PhD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht and CAPHRI Research Institute, University Maastricht.

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

Ankylosing spondylitis (AS), a chronic inflammatory rheumatic disease causing inflammation of the spine and sacroiliac joints, can lead to debilitating pain and stiffness. Characteristic structural bony changes (sclerosis, erosions, bridging, and ankylosis) can be detected on radiographs. But these specific changes often occur late in the course of disease, and their progression is too slow to be measured in short periods of time. Structural changes measured with plain radiography are therefore not very appropriate as outcome measures in short-term clinical trials.

Inflammation of the sacroiliac (SI) joints and the spine, which is the primary abnormality in AS, appears early and fluctuates with shorter time cycles, but cannot be visualized on radiographs. Magnetic resonance imaging (MRI) can be used to visualize inflammation of both SI joints and spine. Its use as an outcome measure in clinical trials seems therefore more rational, provided that the scoring methods for MRI are validated in this context.

During OMERACT 7 in Asilomar, scoring methods for activity in the SI joints were evaluated. It was decided there that the major focus of further research should be scoring of activity of the spine. The conference participants agreed that in light of the available scoring methods for inflammation of the spine in AS emphasis should be on testing aspects of reliability of these different methods, so that a decision about priority of scoring methods could be taken during OMERACT 8[1].

The ASsessments in Ankylosing Spondylitis/OMERACT MRI (ASAS/OMERACT MRI) Working Group decided to evaluate and compare all available scoring methods for inflammation of the spine in AS with respect to feasibility, interobserver reliability, sensitivity to change, and discrimination of MRI.

MATERIALS AND METHODS

Evaluated scoring methods. Three potentially useful scoring methods were identified: the Ankylosing Spondylitis spine Magnetic Resonance Imaging-activity[2] [ASspiMRI-a, acute lesion scores as determined by short-tau inversion recovery (STIR) and gadolinium-enhanced T1 (Gd-DTPA)], the Berlin method[3] (which is a modification of the former with erosions as part of the activity score excluded), and the SPARCC method[4] (Spondyloarthritis Research Consortium of Canada Magnetic Resonance Imaging Index for Assessment of Spinal Inflammation in AS). Scoring systems differed with respect to the MRI sequence required to detect inflammation [T1-weighted turbo-spin echo (TSE) before Gd, same sequence with fat saturation and application of Gd, or STIR], the unit of interest [disco-vertebral units (DVU) divided into quadrants or halves], the number of slices and DVU scored, the qualifications of scoring inflammatory lesions (global grading, extent, intensity), and single versus 3-dimensional evaluation of inflammatory lesions. Consequently the range of the scoring system varied between the methods.

The ASspiMRI-a scoring system. This method uses TSE sequences without fat saturation before Gd, TSE with fat saturation after Gd, and STIR sequences. All 23 DVU of the spine (from C2 to S1), defined as the region between 2 virtual lines through the middle of each vertebra, are scored in a single dimension, which is representing the highest level of inflammation in that particular DVU. Enhancement and bone marrow edema are graded (0-3) for each DVU, with 3 more grades (4-6) if, in addition to the signs of acute inflammation defined for grades 1-3, erosions are visualized, leading to a maximum score of 138 for the entire spine.

The Berlin scoring system. This method is a modification of the ASspiMRI-a system, excluding the score for erosions, so that a DVU can score between 0 and 3, bringing the maximum total score to 69.

The SPARCC scoring system. The entire spine is evaluated for inflammation, but only the 6 most severely affected DVU are scored. This principle was based on a previous study demonstrating that the median number of affected DVU was 3.7[5]. For each detected lesion 3 consecutive sagittal slices are assessed in order to evaluate the extent of inflammation in all 3 dimensions. The presence of an increased STIR signal in each of the quadrants is scored on a dichotomous basis (1: presence; 0: absence) and repeated for each of the 3 consecutive sagittal slices. The presence, on each of the sagittal slices, of a lesion exhibiting high signal intensity (comparable to cerebrospinal fluid) in any DVU is given an additional score of 1. A similar additional score is added in case of a lesion with a continuous depth of $\geq$ 1 cm extending from the endplate. The maximum SPARCC score is 108.

Selection of MRI. Thirty sets of MRI were selected by 2 of us who did not take part in reading (DH, RL) from a randomized controlled trial comparing an active drug [tumor necrosis factor (TNF)-blocking drug, 20 patients] with placebo (10 patients). Sets were selected to cover a wide range of activity at baseline and a wide range of change during followup. One MRI set consisted of paired baseline and post-treatment (week 24) images (T1 before Gd, T1 after Gd, STIR). The readers were unaware of the true time order and of the treatment group. The MRI sets were sent electronically to 10 readers, who were asked to complete a predesigned Excel working sheet following instructions in the form of written guidelines describing the corresponding scoring method. Each reader received the MRI set once, but was asked to rescore the entire set with a different method upon completion of the previous method. Subsequent working sheets with instructions were sent to the readers only if the previous working sheet was returned, and with a time interval of at least 3 weeks in order to minimize recollection. The order of scoring methods by which the readers had to score was randomly defined. Time to score each set (from starting scoring to finishing data input) was also recorded. All readers were members of the ASAS/OMERACT MRI in AS working group, 5 rheumatologists and 4 radiologists.

Training of the readers. Before starting scoring, all readers participated in a training session. During this session the original designers explained the 3 scoring methods and scoring was discussed. Three readers (B1, B2, B3) were experienced in scoring with the ASspiMRI-a scoring system, and therefore by definition with the Berlin method, and 2 (S1, S2) were experienced with the SPARCC method. The remaining 4 readers (N1, N2, N3, N4) did not have experience with any of the scoring methods before training, but are experienced MRI readers. Training images were reviewed and discussed, and scoring guidelines were developed. This training session was organized in an attempt to optimize inter-reader reliability.

Presentation of images. All images were distributed on CD-ROM in DICOM (Digital Imaging and Communications in Medicine) format, enabling com-

patibility with both professional radiological workstations and freely available software packages that can be downloaded from the internet by every participant. Thus, participants of the exercise could use the software environment they were already familiar with. One of us (KGH) developed a manual and assisted in the appropriate installation of the software.

Statistical analysis. Data were aggregated and analyzed by one of us (CL). Feasibility was assessed by comparing mean time-to-score one set of MRI. Inter-reader agreement was determined per scoring method by 2 techniques: smallest detectable change (SDC, calculated from the smallest detectable difference, which is determined from the residual error variance of a repeated measures analysis of variance including all change scores, and is divided by $\sqrt{2}$. The SDC is expressed as an absolute value and as a percentage of the maximum score); and: intraclass correlation coefficient (ICC, single measure, absolute agreement definition) for all readers together and for every possible reader pair separately, both for status scores (baseline and week 24) and change scores.

Sensitivity to change was assessed by calculating Guyatt's effect size per reader and per method on change in scores between baseline and week 24. Guyatt's effect size was calculated by taking the quotient of the mean change score of all patients in the TNF-blocker group and the standard deviation of the change score of all patients in the placebo group[6]. Discrimination between groups (TNF-blocker vs placebo) was compared for the 3 methods using Z-scores from the Mann-Whitney U-test for independent nonparametric observations.

Variance component analysis was conducted using a linear mixed model in order to identify the relevant sources of variability observed in the change scores. To adjust for differences in metric scales across methods, change scores were first standardized on a scale from 0 to 100 (e.g., for the SPARCC method divided by 108 and multiplied by 100). The multivariate model included patient as subject variable, reader, method, the order by which the methods were applied (first, second, or last), and the level of experience with a method as fixed effects. The latter variable was used as a 3-class categorical covariate (experienced in ASspiMRI-a or Berlin, in SPARCC, or in none of them).

RESULTS

Nine of the original ten readers provided completed scoring sheets (30 patients, 2 time points) and these data were used for the analysis. Table 1 shows descriptive results for the scores obtained for each of the 3 evaluated methods. The maximum ASspiMRI-a score observed (55) was at 40% of the scale range (138), the maximum Berlin score (44) at 64% of the scale range, and the maximum SPARCC score (87) at 81% of the scale range. The same picture was seen with regard to change scores. The SPARCC method used the greatest part of the scale range, and the ASspiMRI-a the smallest part. Time to score is evaluated in Table 2. There is an extreme variation in time needed to score for all 3 methods ranging from a few minutes to well over an hour (longest time for the same patient across methods, by 2 readers), but 95% of the patients could be scored within half an hour by all methods. Though the median time needed to score one set (around 10 minutes) is approximately similar for all 3 methods, the time to score for the Berlin method is shorter, with lower mean durations of time [$p = 0.003$ for Berlin vs ASspiMRI-a and $p = 0.001$ for Berlin vs SPARCC (adjusted p values using Bonferroni correction for multiple comparisons)].

Data about inter-reader reliability per method is provided in Table 3. Overall ICC include and evaluate all possible sources of variability among readers, and were highest for SPARCC, lowest for the ASspiMRI-a, and intermediate for the Berlin method. Data for status scores and change scores showed a similar picture across methods.

In order to get an impression about heterogeneity in ICC of all methods, ICC were calculated for every possible reader pair, both for status scores and change scores, and are presented for all 3 methods (Tables 4, 5, and 6). By presenting the data in such a manner, it is easy to discern the level of variability in ICC across all possible reader pairs: reader C, for instance, was found to show markedly different results compared with other readers for the ASspiMRI-a method, and the same pattern can be seen for reader G with the SPARCC method for scoring change, while almost all ICC for reader H

Table 1. Observed status and change scores per method based on the scores of all readers.

	Status Scores (both timepoints)				Change in Scores			
	Minimum	Maximum	Median	SD	Minimum	Maximum	Median	SD
ASspiMRI-a (0–138)	0	63	8	12.9	–36	19	–2	7.81
Berlin (0–69)	0	44	6	8.97	–27	17	–2	5.63
SPARCC (0–108)	0	87	20	21.5	–71	33	–8	17.5

ASspiMRI-a: Ankylosing Spondylitis spine Magnetic Resonance Imaging-activity; SPARCC: Spondyloarthritis Research Consortium of Canada.

Table 2. Time to score one set of images per method.

	Mean	Median	Minimum	Maximum	5th Percentile	95th Percentile
ASspiMRI-a	12 min 44 s	11 min	2 min 5 s	1 h 2 min 31 s	4 min	27 min
Berlin	10 min 16 s	9 min	1 min 8 s	52 min	3 min 3 s	21 min 13 s
SPARCC	13 min 4 s	10 min	34 s	1 h 18 min	2 min 57 s	30 min

ASspiMRI-a: Ankylosing Spondylitis spine Magnetic Resonance Imaging-activity; SPARCC: Spondyloarthritis Research Consortium of Canada.

Table 3. Inter-reader reliability: overall (all 9 readers).

	Intraclass Correlation Coefficient (95% CI)		
	Status Scores (calculated on baseline timepoint)	Status Scores (calculated on Week 24 timepoint)	Change Scores
ASspiMRI-a	0.57 (0.40; 0.73)	0.49 (0.33; 0.66)	0.46 (0.32; 0.63)
Berlin	0.67 (0.49; 0.81)	0.54 (0.37; 0.71)	0.56 (0.42; 0.72)
SPARCC	0.77 (0.66; 0.86)	0.73 (0.61; 0.84)	0.72 (0.61; 0.83)

ASspiMRI-a: Ankylosing Spondylitis spine Magnetic Resonance Imaging-activity; SPARCC: Spondyloarthritis Research Consortium of Canada.

Table 4A. Inter-reader reliability. Method 1: ASspiMRI-a. Intraclass correlation coefficients per reader pair (status scores at baseline).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.39							
Reader N2	0.49	0.57		**Median ICC: 0.60**				
Reader B2	0.65	0.57	0.80					
Reader N3	0.70	0.55	0.49	0.80				
Reader S1	0.73	0.38	0.33	0.58	0.85			
Reader B3	0.68	0.70	0.56	0.73	0.81	0.71		
Reader N4	0.60	0.64	0.83	0.89	0.70	0.49	0.70	
Reader S2	0.34	0.53	0.70	0.60	0.40	0.24	0.43	0.59

Table 4B. Inter-reader reliability. Method 1: ASspiMRI-a. Intraclass correlation coefficients per reader pair (status scores at Week 24).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.60							
Reader N2	0.33	0.35		**Median ICC: 0.53**				
Reader B2	0.67	0.50	0.65					
Reader N3	0.80	0.51	0.27	0.62				
Reader S1	0.70	0.43	0.25	0.52	0.78			
Reader B3	0.68	0.54	0.39	0.60	0.78	0.73		
Reader N3	0.43	0.38	0.76	0.67	0.32	0.36	0.40	
Reader S2	0.40	0.64	0.65	0.53	0.27	0.31	0.36	0.61

Table 4C. Inter-reader reliability. Method 1: ASspiMRI-a. Intraclass correlation coefficients per reader pair (changes in scores).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.25							
Reader N2	0.05	0.17		**Median ICC: 0.47**				
Reader B2	0.56	0.44	0.34					
Reader N3	0.59	0.34	0.22	0.88				
Reader S1	0.49	0.33	0.38	0.51	0.51			
Reader B3	0.55	0.48	0.41	0.75	0.71	0.61		
Reader N4	0.50	0.33	0.36	0.46	0.45	0.86	0.54	
Reader S2	0.38	0.71	0.24	0.64	0.52	0.31	0.61	0.29

Table 5A. Inter-reader reliability. Method 2: Berlin. Intraclass correlation coefficients per reader pair (status scores at baseline).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.73							
Reader N2	0.47	0.57			**Median ICC : 0.71**			
Reader B2	0.68	0.86	0.65					
Reader N3	0.77	0.61	0.36	0.69				
Reader S1	0.71	0.49	0.26	0.47	0.78			
Reader B3	0.79	0.84	0.67	0.89	0.70	0.56		
Reader N4	0.89	0.84	0.61	0.85	0.80	0.66	0.90	
Reader S2	0.63	0.86	0.74	0.88	0.55	0.39	0.87	0.79

Table 5B. Inter-reader reliability. Method 2: Berlin. Intraclass correlation coefficients per reader pair (status scores at Week 24).

	Reader A	Reader B	Reader C	Reader D	Reader E	Reader F	Reader G	Reader H
Reader B	0.67							
Reader C	0.29	0.35			**Median : 0.68**			
Reader D	0.79	0.73	0.42					
Reader E	0.76	0.51	0.23	0.79				
Reader F	0.61	0.42	0.19	0.64	0.83			
Reader G	0.74	0.77	0.39	0.85	0.70	0.63		
Reader H	0.82	0.75	0.38	0.94	0.81	0.67	0.86	
Reader I	0.61	0.73	0.59	0.72	0.46	0.34	0.69	0.70

Table 5C. Inter-reader reliability. Method 2: Berlin. Intraclass correlation coefficients per reader pair (change in scores).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.57							
Reader N2	0.21	0.14			**Median ICC: 0.58**			
Reader B2	0.58	0.68	0.38					
Reader N3	0.81	0.49	0.25	0.67				
Reader S1	0.68	0.42	0.23	0.42	0.65			
Reader B3	0.62	0.58	0.47	0.77	0.62	0.47		
Reader N4	0.80	0.63	0.38	0.71	0.77	0.55	0.74	
Reader S2	0.52	0.75	0.23	0.73	0.55	0.29	0.72	0.66

with the Berlin method were found above the median value. In general, the reader pair analysis shows that the ICC values are consistently higher for the SPARCC method [even the lowest observed ICC (0.47) remains acceptable], and more inconsistent and lower for the ASspiMRI-a and Berlin methods (more variability across reader pairs).

We formally investigated the different sources of variability in change scores by linear mixed modeling, including patients, method, readers, order of method, and level of experience with the method as potential sources of variability. The results of the linear mixed model showed that, in addition to the expected between-patient variability ($p = 0.0002$), the major source of variation in the change scores was, as expected, the method ($p < 0.0001$). Neither the reader ($p = 0.08$) nor the order of the applied method ($p = 0.98$) or the level of experience with any method ($p = 0.08$) contributed significantly to explaining variation in change scores. Another important observation is that the SPARCC method provides similar results for status and change scores, whereas the ASspiMRI-a and Berlin methods show lower inter-reader ICC for change scores compared to status scores.

Apart from ICC we also used the SDC as a reliability statistic. The SDC is the smallest change that can be distinguished from measurement error and can be expressed as the

Table 6A. Inter-reader reliability. Method 3: SPARCC. Intraclass correlation coefficients per reader pair (status scores at baseline).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.68							
Reader N2	0.63	0.78			**Median ICC: 0.76**			
Reader B2	0.73	0.66	0.61					
Reader N3	0.80	0.70	0.71	0.88				
Reader S1	0.71	0.69	0.63	0.92	0.93			
Reader B3	0.65	0.75	0.79	0.71	0.72	0.75		
Reader N4	0.76	0.86	0.76	0.85	0.88	0.87	0.74	
Reader S2	0.69	0.82	0.75	0.87	0.87	0.92	0.79	0.90

Table 6B. Inter-reader reliability. Method 3: SPARCC. Intraclass correlation coefficients per reader pair (status scores at Week 24).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.73							
Reader N2	0.62	0.74			**Median ICC: 0.73**			
Reader B2	0.71	0.68	0.69					
Reader N3	0.73	0.68	0.62	0.92				
Reader S1	0.62	0.63	0.55	0.86	0.90			
Reader B3	0.57	0.69	0.76	0.73	0.67	0.61		
Reader N4	0.71	0.86	0.77	0.86	0.81	0.78	0.84	
Reader S2	0.62	0.77	0.66	0.85	0.82	0.88	0.70	0.86

Table 6C. Inter-reader reliability. Method 3: SPARCC. Intraclass correlation coefficients per reader pair (change in scores).

	Reader N1	Reader B1	Reader N2	Reader B2	Reader N3	Reader S1	Reader B3	Reader N4
Reader B1	0.61							
Reader N2	0.86	0.73			**Median ICC: 0.78**			
Reader B2	0.82	0.59	0.78					
Reader N3	0.88	0.61	0.78	0.73				
Reader S1	0.82	0.66	0.73	0.74	0.93			
Reader B3	0.54	0.42	0.58	0.59	0.55	0.47		
Reader N4	0.81	0.79	0.80	0.76	0.78	0.81	0.52	
Reader S2	0.83	0.85	0.88	0.80	0.82	0.88	0.56	0.86

metric units of the method as well as the percentage of the maximum possible score, in order to improve comparability across methods. SDC can be used as a cutoff to decide if a particular patient has changed more than can be explained by measurement error alone. The SDC in metric units (percentage of the maximum score per method) are 4.1 (3.0%) for the ASspiMRI-a, 2.1 (3.1%) for the Berlin method, and 6.6 (6.1%) for the SPARCC. In order to find an explanation for the apparent paradox that ICC were highest (best) for SPARCC and lowest (worst) for ASspiMRI-a, while SDC were highest (worst) for SPARCC and lowest (best) for ASspiMRI, we visualized all change scores per patient and per method, so that every symbol represents the change score by one reader for one particular patient (Figure 1). The figure shows that in the same set of patients SPARCC uses the highest scoring range and the Berlin method the lowest range. But absolute between-patient variation and absolute within-patient variation is lowest for the Berlin method and highest for the SPARCC method.

Sensitivity to change (Guyatt's effect size) is presented in Table 7. The SPARCC method shows some superiority over Berlin and ASspiMRI-a methods, but the differences were small and sensitivity to change was excellent for all evaluated methods

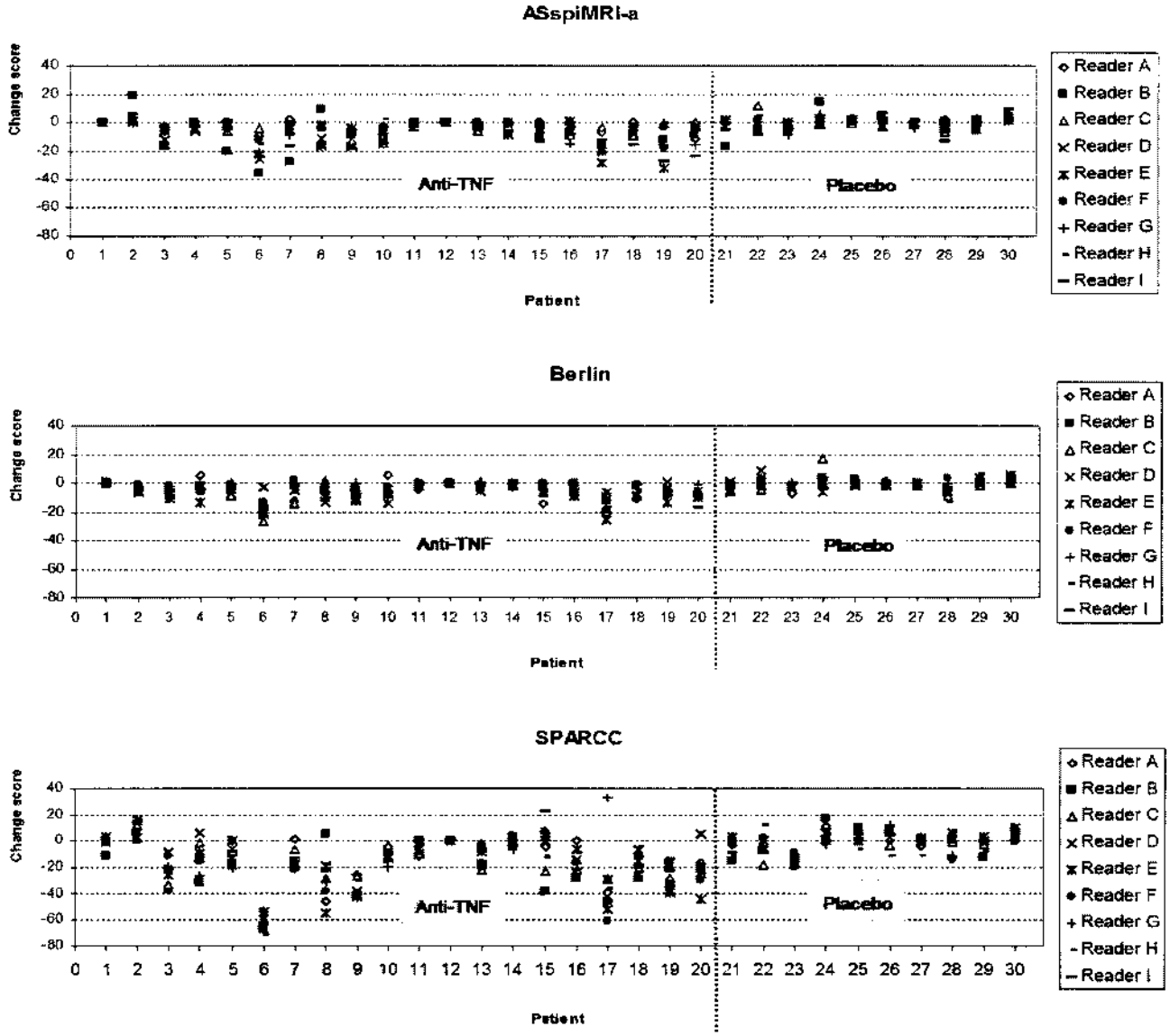

Figure 1. Individual change scores observed in each method.

and for all readers. Especially with the ASspiMRI-a, a few exceptionally high effect sizes (> 4) were found, while the pattern of distribution was more homogeneous for the SPARCC.

Discrimination between groups (TNF-blocker vs placebo) assessed by Z-scores was again very good and comparable among methods, indicating similar between-group discrimination (Table 8).

DISCUSSION

A first conclusion that can be drawn from this study is that we confirmed the feasibility of a worldwide multireader experiment conducted by electronic data dissemination as was already tested for the assessment of SI joint inflammation on MRI with a smaller number of participants[7]. In comparison with the SI joint exercise we have improved methodological quality by using a standardized image format, including training sessions, agreeing on reading rules, and randomizing the order of scoring while forcing a time interval of at least 3 weeks in order to reduce recollection of typical images. It is difficult to judge whether these methodological improvements have really influenced the performance of the readers, but it is clear that in comparison with the SI joints reading experiment, inter-reader ICC of this reading exercise were at a far higher level, both for status and for change scores. More likely explanations are that inflammation of the spine can better be scored than inflammation of the SI joints, that the quality of the films of the spine was far better than the quality of the films of the SI joints, and that there was much more active inflammation in the spine compared to the SI joints. A second advantage of this reading exercise in comparison with the previous SI joint exercise is that more detailed information on discrimination (Guyatt's effect sizes

Table 7. Sensitivity to change: Guyatt's effect size per method per reader.

	N1	B1	N2	B2	N3	S1	B3	N4	S2	Median
ASspiMRI-a	1.96	0.99	0.95	5.66	4.1	2.40	1.76	0.97	1.69	1.76
Berlin	1.48	0.98	1.08	2.74	2.16	2.74	1.24	1.83	1.62	1.62
SPARCC	1.71	1.75	2.30	2.20	3.06	2.48	2.07	1.60	2.06	2.07

Table 8. Discrimination ability: Z score (Mann-Whitney test comparing change score between patients treated with anti-TNF drug vs placebo) per method per reader.

	N1	B1	N2	B2	N3	S1	B3	N4	S2	Median
ASspiMRI-a	-2.578	-1.631	-2.474	-3.438	-2.867	-3.245	-3.006	-2.461	-2.511	-2.578
Berlin	-2.323	-1.986	-2.475	-3.133	-2.945	-2.041	-1.941	-2.903	-2.380	-2.380
SPARCC	-2.556	-2.510	-2.710	-2.605	-3.180	-2.733	-2.400	-1.982	-2.472	-2.556

and between-group Z-scores) could be determined, and that time-to-score was assessed as a major determinant of feasibility. As a consequence, we consider the quality of this reading experiment on spinal MR images as improved compared to that of the previous experiment on SI joints, and the conclusions derived from this experiment should therefore have a greater influence.

What are the main conclusions about the content of the experiment? First, the SPARCC method outweighed the ASspiMRI-a and Berlin methods with respect to inter-reader ICC, particularly with regard to change scores, but also with regard to status scores. The distinctive difference is that — when looking at ICC of different reader pairs — even the worst reader pair obtained an acceptable level of agreement with the SPARCC method (for change scores 0.42), while agreement for the worst reader pair was completely lacking with the other methods (0.05 and 0.14). This homogeneity was somewhat surprising, because the SPARCC method yielded the highest variation in absolute change scores (scale range), and the SPARCC method implies an additional source of variability: the choice of the 6 levels with highest inflammatory activity.

There is a technical problem that jeopardizes the comparison of ICC across methods, namely that such a comparison assumes a similar between-patient variability (e.g., 2 readers score the same set of patients with the same method and consequently use approximately the same range of the scale). Figure 1 clearly shows that between-patient variability (scale range) is very different across methods, even after standardization. Interestingly, absolute within-patient variation (by readers) was far lower for Berlin compared to SPARCC, pointing to better agreement with Berlin. This difference in absolute within-patient variation is reflected by differences in the SDC that are higher (worse) for SPARCC compared to the other methods, pointing to worse absolute agreement. However, ICC are higher for SPARCC, pointing to better relative agreement. The SDC quantifies the level of absolute variation in the data. The ICC estimates the proportion of variance in the data that is due to differences between the subjects rather than differences between the readers, and as such reflects the concept of the signal-to-noise ratio. This multireader experiment elegantly demonstrates that any conclusion about reliability is fundamentally dependent on the choice of the reliability statistic. We have shown here that the choice of a statistic that relies on "relative agreement" (ICC) may give opposite results compared to a statistic that relies on "absolute agreement" (SDC). This paradox has been shown previously in the literature, and there is no unanimity about the best reliability statistic under certain circumstances[8]. We refrain from an exhaustive discussion about reliability statistics here, but it is to be noted that an ICC value is biased towards high coefficients (close to 1, which means optimal agreement) if the data vary over a wide range (such as SPARCC). The SDC, however, is biased towards smaller values (closer to zero, which means perfect agreement) if the data vary over a narrow range of values (such as the Berlin method)[8]. It is therefore impossible to conclude whether the higher ICC overestimate reliability of SPARCC, or that the lower SDC overestimate reliability of the ASspiMRI-a and the Berlin method. However, the homogeneity of reader-pair ICC for the SPARCC may introduce another advantage for this method in that it is less dependent on the choice of the particular reader pair.

A number of characteristics of the SPARCC method may theoretically point to reduced inter-reader variability compared to other methods: The change score pertains to only 6 levels with the most severe inflammation, whereas the ASspiMRI-a and Berlin methods, that score the entire spine, include a majority of levels with no or dubious inflammation, that can easily be a source of noise in scores. Another advantage of the SPARCC method, which may limit inter-reader variability, is the binomial answering modality: Inflammation in a quadrant is either present or absent, and additional points for depth and/or intensity are clearly defined. The ASspiMRI-a and Berlin methods embark on graded answering modalities, which leave room for interpretation differences, and the ASspiMRI-a method requires the separate interpretation of

erosions, which is another domain. The last characteristic of the SPARCC method that may constrain inter-reader variability is the preferential choice for one sequence (STIR) to score inflammation, whereas the ASspiMRI-a method is defined for both STIR and post-Gd sequences without clear guidance for a preferential sequence. However, it was shown that the use of STIR sequences alone for the ASspiMRI-a is sufficient in the setting of clinical trials[9,10].

Two other characteristics of discrimination were studied in this experiment. Sensitivity to change and between-group discrimination were approximately similar across methods. It is important to mention here that sensitivity to change was actually very good for all 3 methods, and several readers achieved extremely high effect sizes. Such high effect sizes according to Guyatt's method can only be reached if the effect (change) in the active intervention group is very good and the response in the placebo group very homogeneous (low standard deviation). It is also important to mention here that this experiment was not very appropriate to investigate between-group discrimination because the number of patients was far too low (unlike effect sizes, between-group Mann-Whitney Z-scores are sensitive to patient numbers), and the patients were not a representative sample from the randomized trial; they were selected on the basis of superior imaging quality and extent of inflammation, while assuring a graded representation of the entire spectrum of inflammatory changes. Nevertheless, all 3 methods were tested under the same artificial circumstances, and differences in Z-scores between methods can be interpreted as long as the absolute value of the Z-score is not given any importance. Given these limitations, the Z-scores of all methods were very comparable. Feasibility was addressed by assessing time-to-score. All methods showed a similar median time-to-score. A possible explanation is that the additional time needed to make an adequate choice of levels in the SPARCC method is effaced by the fact that 17 of the 23 potential levels (that should be scored in the ASspiMRI-a and Berlin method) could be ignored.

So, in summary, testing different aspects of discrimination and feasibility showed that the SPARCC method consistently shows higher ICC and increased consistency in ICC values between different reader pairs than the ASspiMRI-a method and the Berlin method, but the SDC are smaller for ASspiMRI-a and Berlin, making a correct judgment about the most reliable method difficult. Sensitivity to change, between-group discrimination, and feasibility of the 3 methods were comparable, and at a more than acceptable level.

With regard to the Truth aspect of the OMERACT filter, the limited number of vertebrae to score in the SPARCC can be considered a disadvantage in terms of generalizability, in comparison with both other methods that score the entire spine, and may give a better representation of spinal inflammation. On the other hand, the SPARCC better reflects the advantage of MRI that allows the evaluation of lesions in more than one dimension. An important difference between the ASspiMRI-a method and the Berlin method is that the former includes erosions as an activity criterion whereas the latter does not weigh erosions as such. A conclusion from our study that does not show important differences in psychometric properties between ASspiMRI-a and Berlin could be that erosions are not very important. But with regard to the Truth aspect of the OMERACT filter, the contribution of erosions is still unclear. One truth aspect that deserves attention in the future — irrespective of the chosen method — is the correlation of MRI activity with clinical variables such as pain and function. Another important aspect is predictive validity, or: does MRI activity predict function loss or structural damage?

Taking all these arguments into account, it is difficult to prioritize one of the 3 methods for scoring inflammation on the basis of our multireader experiment. The SPARCC method may have advantages in terms of reliability, especially since it demonstrates more consistency in this regard, whereas the ASspiMRI-a and Berlin method provide a better overall representation of inflammation of the spine.

REFERENCES

1. van der Heijde DM, Landewe RB, Hermann KG, et al. Application of the OMERACT filter to scoring methods for magnetic resonance imaging of the sacroiliac joints and the spine. Recommendations for a research agenda at OMERACT 7. J Rheumatol 2005;32:2042-7.
2. Braun J, Baraliakos X, Golder W, et al. Magnetic resonance imaging examinations of the spine in patients with ankylosing spondylitis, before and after successful therapy with infliximab: Evaluation of a new scoring system. Arthritis Rheum 2003;48:1126-36.
3. Rudwaleit M, Schwarzlose S, Listing J, Brandt J, Braun J, Sieper J. Is there a place for magnetic resonance imaging (MRI) in predicting a major clinical response (BASDAI 50%) to TNF alpha blockers in ankylosing spondylitis? [abstract] Arthritis Rheum 2003;50:S211.
4. 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.
5. Braun J, Baraliakos X, Golder W, et al. Analysing chronic spinal changes in ankylosing spondylitis: a systematic comparison of conventional x rays with magnetic resonance imaging using established and new scoring systems. Ann Rheum Dis 2004;63:1046-55.
6. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.
7. Landewe RB, Hermann KG, van der Heijde DM, et al. Scoring sacroiliac joints by magnetic resonance imaging. A multiple-reader reliability experiment. J Rheumatol 2005;32:2050-5.
8. Lassere M, McQueen F, Ostergaard M, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 3: an international multicenter reliability study using the RA-MRI Score. J Rheumatol 2003;30:1366-75.
9. Baraliakos X, Hermann KG, Landewe R, et al. Assessment of acute spinal inflammation in patients with ankylosing spondylitis by magnetic resonance imaging: a comparison between contrast enhanced T1 and short tau inversion recovery (STIR) sequences. Ann Rheum Dis 2005;64:1141-4.
10. Hermann KG, Landewe RB, Braun J, van der Heijde DM. Magnetic resonance imaging of inflammatory lesions in the spine in ankylosing spondylitis clinical trials: is paramagnetic contrast medium necessary? J Rheumatol 2005;32:2056-60.