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

DÉSIRÉE M.F.M. VAN DER HEIJDE, ROBERT B.M. LANDEWÉ, KAY-GEERT A. HERMANN, ANNE-GRETHE JURIK, WALTER P. MAKSYMOWYCH, MARTIN RUDWALEIT, PHILIP J. O'CONNOR, JÜRGEN BRAUN, and the ASAS/OMERACT MRI in AS Working Group

ABSTRACT. Magnetic resolution imaging (MRI) is a promising tool in the assessment of inflammation and structural damage in clinical trials in ankylosing spondylitis (AS). The ASAS/OMERACT MRI in AS working group, a collaborative initiative of rheumatologists and musculoskeletal radiologists with a special interest in this field, collected data on all available scoring methods for both sacroiliac (SI) joints and spine, and tested them with respect to the OMERACT filter. These data were presented together with the technical specifications of all methods at the OMERACT 7 conference. In addition, the results of 2 separate experiments on the inter-reader reliability of scoring methods to assess activity in SI joints, and on the comparison of STIR sequence versus T1 post-gadolinium (Gd) sequence for the spine, were presented. Thereafter, 8 groups discussed these data and proposed a research agenda, each on a different topic. This information was reported back to all participants and a prioritized research agenda was compiled by voting. Research on scoring methods for assessing disease activity, in both the spine and SI joints, was considered most important. Research on assessing structural damage was considered less important. The specific process and results of this initiative are discussed. (J Rheumatol 2005;32:2042–7)

Key Indexing Terms:
IMAGING
SCORING METHODS
ANKYLOSING SPONDYLITIS
MAGNETIC RESONANCE IMAGING

From the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht, Maastricht, The Netherlands; Department of Radiology, Charité Medical School, Berlin, Germany; Department of Radiology, Aarhus University Hospital, Aarhus, Denmark; Department of Medicine, University of Alberta, Edmonton, Canada; Department of Radiology, General Infirmary at Leeds, Leeds, UK; and the Rheumazentrum Ruhrgebiet, Herne, Germany.

D.M.F.M. van der Heijde, MD, PhD, Professor of Rheumatology; R.B.M. Landewé, MD, PhD, Associate Professor of Rheumatology, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht; K-G.A. Hermann, MD, Radiologist, Department of Radiology, Charité Medical School; A-G. Jurik, MD, Radiologist, Department of Radiology, Aarhus University Hospital; W.P. Maksymowych, FRCPC, Professor of Medicine, Department of Medicine, University of Alberta; M. Rudwaleit, MD, Rheumatologist, Department of Rheumatology, Charité Medical School, Campus Benjamin Franklin; P.J. O'Connor, MD, Radiologist, Department of Radiology, Leeds General Infirmary; J. Braun, MD, Professor of Rheumatology, Rheumazentrum Ruhrgebiet.

Members of the Assessment in Ankylosing Spondylitis/OMERACT MRI 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, Rob Lambert, Robert Landewé, Walter P. 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 Prof. D.M.F.M. van der Heijde, University Hospital Maastricht, Department of Internal Medicine, Division of Rheumatology, PO Box 5800, 6202 AZ Maastricht, The Netherlands.

The international ASessment in Ankylosing Spondylitis (ASAS) working group has recognized imaging as an important domain in ankylosing spondylitis (AS) and recommends further development for use in clinical trials[1]. Magnetic resonance imaging (MRI) is becoming an important tool for assessment of inflammation and structural damage in ankylosing spondylitis (AS). Both the sacroiliac (SI) joints and the spine are sites of particular interest in the assessment of AS. Several groups worldwide have been involved in research on this topic, including the development of scoring systems for acute and chronic changes in SI joints and spine. One of the prerequisites for instruments to be useful is that they must have passed the OMERACT filter successfully with respect to truth, discrimination, and feasibility[2]. In order to further develop MRI applications for clinical trials in AS, the ASAS/OMERACT working group for MRI in AS was established. This working group is a followup initiative from MISS (MR Imaging of Seronegative Spondylarthropathy), which was started earlier by the Leeds UK group. Rheumatologists and muskuloskeletal radiologists that had proven interest (research and publications) in MR imaging of the SI joints and the spine were invited to meet in Orlando in November 2003. During that meeting,

Downloaded on April 24, 2024 from www.jrheum.org

the potential importance of MRI in AS clinical trials was broadly recognized, and it was agreed to perform a collaborative action under the umbrella of ASAS and OMERACT to further develop MRI of the SI joints and the spine as a measurement instrument in AS clinical trials. The ultimate goal is to have one valid scoring system (for SI joints and for spine, for inflammatory and for structural changes) accepted by the rheumatological and radiological community involved in the assessment of AS. In view of the available literature, the group decided to focus first on measuring inflammatory activity of the SI joints. The group met a second time in January 2004 in Bethesda and communicated extensively by E-mail in preparation for a module on MRI in AS held at OMERACT 7. As part of the preparation for the OMERACT module, an exercise on comparing inter-reader reliability and sensitivity to change of scoring methods for inflammatory changes in the SI joints was performed by this group. This article describes the process followed and the decisions taken during the OMERACT 7 conference.

Process

The aim of the OMERACT module was to define a prioritized research agenda for the next 2 years for further development and validation of MRI scoring methods for inflammatory and structural changes of the SI joints and the spine. The module started with a plenary session presenting general background information on the use of MRI in AS (K.G. Hermann), followed by a literature review on the various aspects of the OMERACT filter for the existing scoring systems (J. Braun), and the results of 2 research projects (one on inflammatory changes in the SI joints, and another on the comparison of short-tau inversion recovery (STIR) and post-gadolinium (Gd) sequences on the assessment of inflammation of the spine) (R.B.M. Landewé). These latter 2 research projects are published in these proceedings[3,4]. After the plenary session, 8 discussion groups were formed and assigned various tasks. One rapporteur per group reported back in a second plenary session, which was followed by voting on specific questions by all participants developing the research agenda.

Overview of the Available Scoring Methods

Based on a literature review, combined with unpublished information from the members of the ASAS/OMERACT MRI in AS working group, tables were compiled with the available data on intra- and inter-reader reliability, change over time, and discrimination between patients. This was done separately for the SI joints and for the spine, and for inflammatory and structural changes. Six methods were identified for assessing inflammatory lesions in SI joints: the MISS scoring system, the Leeds scoring system, the Aarhus scoring system, the Spondyloarthritis Research Consortium Canada (SPARCC) scoring system, and 2 initiatives from Berlin by Sieper and Rudwaleit, and Hermann and Bollow[5-10]. Only the system developed in Aarhus has been published as a full article. The MISS and SPARCC scoring systems have been published in abstract form, the Berlin Hermann/Bollow system as a proposal as yet unvalidated, and the Leeds and Berlin Sieper/Rudwaleit systems have not yet been published.

The details of acquisition and method of scoring and grading are presented in Table 1. Some methods use Gd enhancement, while others use STIR only. Scoring ranges from a global score for the entire joint to a detailed scoring of several slices in quadrants. Table 2 provides data on reliability of the methods, which were mostly unpublished. Change over time and discrimination between patients was almost never investigated. Intra-reader reliability (based on kappa statistics) was mostly good to excellent, but the inter-reader reliability was only poor to moderate, except for the SPARCC method, which was very good. In general, all this information is based on small numbers of images and readers, and was only obtained from the centers where the methods were developed.

Four scoring methods for assessing structural changes of the SI joints were proposed: MISS, Leeds, Aarhus, and Berlin (Hermann/Bollow)[5,6,9]. Technical details are presented in Table 3. Only the method developed by the Aarhus group had data on reliability (presented in Table 4): intra-reader reliability was good, but only moderate inter-reader reliability was found.

Four scoring methods (SPARCC, Leeds, Berlin Sieper/Rudwaleit, and ASspiMRI-a) have been proposed for assessing active lesions in the spine[10,11]. The technical specifications of the methods are presented in Table 5 and the data on reliability in Table 6. Only the ASspiMRI-a score uses Gd enhancement as a standard. The Berlin method is based on the ASspiMRI-a, with the modification that Gd enhancement is not obligatory, and that erosions are not included in the activity score. These 2 methods score all vertebrae from C2 to S1, while the Leeds system includes only the lumbar vertebrae, and the SPARCC system scores the 6 most severely affected disco-vertebral segments. Only for the SPARCC and ASspiMRI-a methods was complete information on reliability provided. For both methods, the intra- and inter-reader reliability was good to excellent. In addition, sensitivity to change and discrimination between patient groups has been shown by the ASspiMRI-a, and sensitivity to change with the SPARCC.

Finally, 2 methods have been proposed for scoring structural damage lesions: the Leeds and the ASspiMRI-c system (Table 7)[11]. Only for the latter were data available on intra-reader and inter-reader reliability, showing good and poor to moderate reliability, respectively (Table 8).

Group Discussions

In total, 8 groups were formed and assigned specific topics;

Downloaded on April 24, 2024 from www.jrheum.org

Table 1. Technical specifications of scoring methods assessing activity in SI joints.

Scoring Method	Sequences	Orientation	Slice Thickness	Score Per...	Segments	Grades	Range
MISS[5]	T1 pre-Gd, T2 fat-suppressed	Coronal oblique		Halves	2 SI joints	Extent (0–3), intensity (0–2), global activity (0–2)	0–12 0–8 0–8
Leeds	T1 turbo spin-echo, T2 SPIR FS, T1 FFE SPIR post-Gd	Coronal oblique		Quadrant	2 SI joints	Intraarticular activity (0–3), subchondral activity (0–3) Change between scans: resolution, improvement, no change, new lesion	0–12 0–12
Aarhus[6]	STIR, T1, T1 FS before and after Gd, T2	Coronal oblique Axial oblique	4 mm	Quadrant for osseous portion and 2 joint spaces (cartilaginous and ligamentous)	2 SI joints	Bone marrow edema (0–3) Gd enhancement bone marrow (0–3), Gd enhancement joint space (0–3)	0–60
Berlin (Sieper/Rudwaleit)	STIR (Gd if STIR not available)			Halves	2 SI joints	Bone marrow edema (0–3)	0–12
SPARCC[7]	T1 SE (as reference), T2 STIR	Coronal oblique	3–4 mm (total 12 slices)	Quadrants	Slices 4–9 of 2 SI joints	Inflammation 0–1 per quadrant per slice; per joint extra point for intensity and for depth	0–72
Berlin (Hermann/Bollow)[8]	T1, STIR, post-Gd, T2	Coronal oblique	4 mm	Quadrants	2 SI joints	Increased signal in joint space plus bone marrow edema (0–4)	0–32

FFE: fast field echo; FS: fat suppression; Gd: gadolinium; SE: spin echo; SPIR: spectral presaturation with inversion recovery; STIR: short-tau inversion recovery.

Table 2. Data on discrimination of scoring methods assessing activity in SI joints.

Scoring Methods SI Activity—Reliability	No. of Patients	Intra-Reader	Inter-Reader	Change Over Time	Between Patients
MISS[5]	N = 20	—	N = 8, per quadrant ICC	—	—
Extent			0.19–0.44		
Intensity			0.30–0.51		
Global			0.19–0.45		
Leeds	?	Kappa 0.44–0.83	Kappa 0.44–0.83	Present	—
Aarhus[6]	N = 41	Kappa 0.84–1.00	N = 2, kappa	—	—
Bone marrow edema			0.49		
Enhancement bone			0.47		
Enhancement joint space			0.64		
Activity score			0.29		
Berlin (Sieper/Rudwaleit)	—	—	—	—	—
SPARCC[7]	N = 22	Kappa 0.90–0.98	N = 3, kappa 0.89	ES = 0.1–0.22 (3 joints fused)	—
Berlin (Hermann/Bollow)[8]	—	—	—	—	—

ES: effect size.

each group had a discussion leader and a rapporteur. Four groups discussed scoring methods for inflammatory lesions; 2 discussed methods for the SI joints and 2 for the spine; 2 groups discussed scoring methods for structural changes, one for the SI joints and one for the spine.

For their specific assignment each group discussed available data on reliability, sensitivity to change, face validity, and what data are lacking. In addition, the groups discussing the spine also considered whether zygo-apophyseal joints and ligaments should be included in a scoring method. The final 2 groups discussed the results of 2 experiments: the group discussing the SI scoring experiment on activity was asked to prioritize the methods based on the inter-reader data and based on the sensitivity to change data. The group discussing the post-Gd versus STIR experiment considered whether STIR images are sufficient to assess activity of the spine, and whether these data are sufficient to generalize to the SI joints.

Downloaded on April 24, 2024 from www.jrheum.org

Table 3. Technical specifications of scoring methods assessing structural changes in SI joints.

Scoring Method	Sequences	Orientation	Slice Thickness	Score Per...	Segments	Grades	Range
MISS[5]	T1 pre-Gd, T2 fat-suppressed	Coronal oblique		Halves	2 SI joints	Global chronicity NY criteria (0–4)	0–8
Leeds	T1 turbo spin-echo, T2 SPIR FS, T1 FFE SPIR post-Gd	Coronal oblique		Quadrant	2 SI joints	Subchondral sclerosis (0–3), ankylosis (0-3) 0 = absent, 1 = mild 0–25%, 2 = moderate 25–75%, 3 = severe > 75%; Change between scans: resolution, improvement, no change, new lesion	0–12 0–12
Aarhus[6]	STIR, T1, T1 FS before and after Gd, T2	Coronal oblique Axial oblique		Quadrant for osseous portion and 2 joint spaces (cartilaginous and ligamentous)	2 SI joints	Erosion (0–3), sclerosis (0–3), joint space width (0–3) Separate: global (0–4)	0–60 0–4 0–4
Berlin (Hermann/Bollow)[9]	T1, STIR, post-Gd, T2	Coronal oblique	4 mm	Whole joint	2 SI joints	Global chronicity similar to NY criteria (0–4)	0–4 per joint

SI: sacroiliac. For other definitions, see note to Table 1.

Table 4. Data on discrimination of scoring methods assessing chronic changes in SI joints.

Scoring Methods SI Chronicity—Reliability	No. of Patients	Intra-Reader	Inter-Reader	Change Over Time	Between Patients
Aarhus[4]	N = 41	Kappa 0.84–1.00	N = 2, kappa	—	—
Erosion			0.54		
Sclerosis			0.41		
Joint width			0.18		
Joint score			0.42		

Table 5. Technical specifications of scoring methods assessing activity in the spine.

Scoring Method	Sequences	Orientation	Slice Thickness	Score Per...	Segments	Grades	Range
SPARCC[10]	T1 SE (as reference), STIR	Sagittal	3–4 mm (total 12 slices)	Disco-vertebral unit divided into 4 quadrants	6 units showing the most apparent lesions on STIR; for each lesion 3 consecutive sagittal slices are assessed	12 for edema for each disco-vertebral unit, extra point for intensity and depth; total per unit (3 slices) 18; grand total for 6 units 108	0–108
Leeds	T2 SPIR	Sagittal		Vertebral body, spinous processus, facetal joints, paraspinal soft tissue	5 lumbar vertebrae	No. of lesions	
Berlin (Sieper/Rudwaleit)	STIR (Gd only if STIR not available)	Sagittal		Vertebral unit	23 vertebral units (C2/C3—L5/S1)	Bone marrow edema (0–3)	0–69
ASspiMRI-a[11]	Gd, STIR	Sagittal		Vertebral unit	23 vertebral units (C2—S1)	0–6	0–138

For definitions, see note to Table 1.

The rapporteurs reported back to the general audience. The conclusion from all groups was that for all the methods there was too little information on reliability and sensitivity to change, so ranking of methods was not possible. Another general comment was that scoring methods for inflammatory changes were more useful than those for assessing structural damage.

The group discussing the SI experiment commented that the presented interclass correlation coefficients were good for all methods, given the lack of training and the involve-

Downloaded on April 24, 2024 from www.jrheum.org

Table 6. Data on discrimination of scoring methods assessing activity in the spine.

Scoring Methods SI Spine Activity—Reliability	No. of Patients	Intra-Reader	Inter-Reader	Change Over Time	Between Patients
SPARCC[10]	N = 22	N = 11 kappa 0.93–0.98 (STIR)	N = 22 kappa 0.80 (STIR)	ES = 0.73–0.82	—
Leeds No. of lesions (change)	—	—	—	Present	—
Berlin (Sieper/Rudwaleit)	—	—	—	—	—
ASspiMRI-a[11]	N = 20	Gd Var comp 8.1% / STIR Var comp 5.0%	N = 2 Gd Var comp 6.8% / STIR Var comp 15.0%	Present	Present

For definitions see note to Table 1.

Table 7. Technical specifications of scoring methods assessing structural changes in the spine.

Scoring Method	Sequences	Orientation	Slice Thickness	Score Per...	Segments	Grades	Range
Leeds	T2 SPIR	Sagittal					
ASspiMRI-c[11]	T1, T2	Sagittal		Vertebral unit	23 vertebral units (CS2—S1)	0–6	0–138

For definitions see note to Table 1.

Table 8. Data on discrimination of scoring methods assessing structural changes in the spine.

Scoring Methods Spine Structural Changes—Reliability	No. of Patients	Intra-Reader	Inter-Reader	Change Over Time	Between Patients
ASspiMRI-c[11]	N = 20	Variance component 21.7%	Variance component 47.6%	Present	—

Table 9. Research agenda.

Scoring methods to assess activity in SI joints
- Develop a validation protocol
- Every developer of a scoring method will further validate their method according to a consensus validation protocol
- Thereafter cross-validation of every method by other groups

Scoring methods to assess activity in spine
- Is bone marrow edema specific for inflammation in AS?
- Will edema result in erosions?
- Usefulness of including zygo-apophyseal joints in a scoring method
- Which features should be included in a scoring method?
- Which/how many vertebrae should be assessed?
- Comparison of the 3 available scoring methods
- Comparison of relative contribution from STIR and post-Gd images

ment of 7 readers. Based on the data, no ranking of the methods could be determined at this time. The recommendation by the groups on assessing activity in SI joints was to develop a validation protocol by consensus, and to ask developers to apply this validation protocol to their own method. Thereafter, in a second exercise, the comparative reliability of the methods could be assessed by readers scoring different methods.

The group that discussed scoring methods for chronic changes concluded that too little information was available to draw conclusions. Moreover, they expressed special concerns about the lack of appropriate definitions for abnormalities.

The usefulness of MRI films for assessing structural damage was questioned; this topic was therefore given low priority for further research.

The groups discussing scoring methods for assessing

Downloaded on April 24, 2024 from www.jrheum.org

activity in the spine stressed as an important issue the degree of specificity of bone marrow edema for inflammation in AS. Second, they recommended further studies to determine whether edema is followed by erosions. Concerning face validity, the group found advantages in the ASspiMRI-a method as this includes the entire spine. On the other hand, the Leeds method captures more features. Scoring of the zygo-apophyseal joints may provide important information, but on the other hand scoring these joints may be very difficult. Therefore, the usefulness (how much extra information is gained at the cost of extra noise) of this approach needs to be investigated. Also to be further addressed: whether erosions should be included in scoring activity, as this might be a sign of structural change rather than inflammation.

In more general terms, the groups asked for further studies on the relative contribution of the various features. The group discussing the ASspiMRI-c scoring method for assessing structural changes of the damage decided that the method has sufficient face validity. However, further research needs to be done with regard to the OMERACT filter. With respect to this, a clear definition should be given for the specific features, and comparison with other methods assessing structural damage (radiographs, computerized tomographic scan) should be undertaken. The group discussing the STIR/post-Gd experiment concluded that there was comparable performance of both methods with respect to reliability and sensitivity to change. Because of fewer costs, better standardization, and less time to scan, the STIR sequence may be preferred. However, one should keep in mind that this preference is based on a relatively small number of patients, that Gd-enhanced images may occasionally facilitate delineation of structural changes, and that there might be discordance between STIR and post-Gd at the individual patient level.

Voting

After reporting to the plenary audience, there was voting on various questions in order to prioritize the research agenda. Fifty-two percent of participants indicated MR images of the SI joints might be a useful outcome measure for assessing activity in a clinical trial on drug efficacy, 9% disagreed, and 31% judged there were not enough data. On the same question regarding assessment of chronic changes, 30% answered yes, 31% no, and 36% not enough data. For assessing activity in the spine, 74% of the audience voted yes, only 2% no, and 22% not enough data. A similar divided opinion was recorded for assessing chronic changes in the spine as well as for assessing SI joints: 30% yes, 21% no, and 48% not enough data. The majority response to whether there were sufficient data to decide that both STIR and post-Gd sequences or only one of them should be routinely performed in clinical trials for assessing activity in the spine was "no," which puts this research question back on the research agenda.

During the final summary session, a large majority of participants confirmed by voting that it was more important to examine scoring methods for assessing inflammatory rather than structural changes, and that the spine was more promising than the SI joints.

Conclusion

MRI as an outcome measure in clinical trials in AS is a promising tool. Still, a lot of work needs to be done on further validating the various methods, which will hopefully result in one valid, standardized method for use in clinical trials. There was general agreement that developing methods for assessing active inflammation takes priority over methods to assess structural changes. The ASAS/MRI in AS working group proved to be an active and productive group of collaborators, willing to develop this area further.

REFERENCES

1. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. J Rheumatol 1997;24:2225-9.
2. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology [editorial]. J Rheumatol 1998;25:198-9.
3. Landewe 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.
4. Hermann K-GA, Landewé RBM, Braun J, et al. Magnetic resonance imaging of inflammatory lesions in the spine in ankylosing spondylitis clinical trials: Is paramagnetic contrast medium necessary? J Rheumatol 2005;32:2056.
5. 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.
6. 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.
7. 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 1:76.
8. Hermann K-GA, Braun J, Fischer T, Reisshauer BH, Bollow M. Magnetic resonance imaging of sacroiliitis: anatomy, histological pathology, MR-morphology, and grading [German]. Radiologe 2004;44:217-28.
9. Bollow M, Braun J, Taupitz M, et al. CT-guided intraarticular corticosteroid injection into the sacroiliac joints in patients with spondyloarthropathy: indication and follow-up with contrast-enhanced MRI. J Comput Assist Tomogr 1996;20:512-21
10. 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 spinal inflammation in spondyloarthritis [abstract]. Ann Rheum Dis 2004;63 Suppl 1:90.
11. 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.

Downloaded on April 24, 2024 from www.jrheum.org