Scoring with the Berlin MRI method for assessment of spinal inflammatory activity in patients with ankylosing spondylitis: a calibration exercise among rheumatologists

L. Carmona[1], A. Sellas[2], C. Rodríguez-Lozano[3], X. Juanola[4], J.F. García Llorente[5],
J.L. Fernández Sueiro[6], L.F. Linares[7], M.C. de Castro[8], M. Moreno[9], P. Zarco[10],
R. Ariza[11], X. Baraliakos[12], E. de Miguel[13]

[1]Institute for Musculoskeletal Health, Madrid; [2]Rheumatology Service, Hospital Vall d'Hebron, Barcelona; [3]Rheumatology Service, Hospital Doctor Negrín, Las Palmas de Gran Canaria; [4]Rheumatology Service, Corporacio Sanitària i Universitària Parc Taulí, Sabadell; [5]Rheumatology Service, Hospital de Basurto, Bilbao; [6]Rheumatology Service, CH Universitario A Coruña, A Coruña; [7]Rheumatology Service, Hospital Virgen de la Arrixaca, Murcia; [8]Rheumatology Service, Hospital Reina Sofía, Córdoba; [9]Rheumatology Service, Hospital Universitari de Bellvitge, Barcelona; [10]Rheumatology Service, Fundación Hospital de Alcorcón, Madrid; [11]Rheumatology Service, Hospital Universitario Virgen Macarena, Seville, Spain; [12]Rheumazentrum Ruhrgebiet, Ruhr-University Bochum, Germany; [13]Rheumatology Department, Hospital Universitario de la Paz, Madrid, Spain.

Abstract

Objective

To test the reliability of the Berlin MRI scoring method and the effect of a calibration exercise on the score's reliability among untrained readers in MRI examinations of patients with established ankylosing spondylitis (AS).

Methods

Eleven rheumatologists read blinded images of 20 AS patients before and after a two-day workshop on the Berlin MRI scoring method. Reliability (intra- and inter-reader) and concordance with the expert (all measured by intraclass correlation coefficient (ICC)) were compared before and after 2 weeks of the training. Feasibility in terms of time and difficulty was also measured.

Results

The mean Berlin score increased from (mean ± standard deviation) 5.04 ± 6.41 before to 6.40±7.08 after the calibration exercise ($p<0.01$). Inter-reader ICC decreased from 0.83 (95% CI: 0.75–0.93) to 0.78 (95% CI: 0.66–0.90), and intra-reader ICC from 0.89 (95% CI: 0.84–0.94) to 0.87 (95% CI: 0.82–0.92). Agreement with an experienced reader improved after the calibration exercise, with ICC = 0.59 (95% CI 0.45–0.76) before vs. ICC = 0.65 (95% CI 0.50–0.80) after training.

Conclusion

The Berlin method is a reliable scoring method for assessment of spinal inflammatory activity by using MRI in patients with AS, even in the hands of inexperienced readers. A calibration exercise can improve feasibility and sensitivity of the scoring method.

Key words

education, reproducibility of results, ASspiMRI, reliability, spondyloarthritis, validation studies

Clinical and Experimental Rheumatology 2013; **31**: 883-888.

Loreto Carmona, MD, PhD
Agustí Sellas, MD
Carlos Rodríguez-Lozano, MD
Xavier Juanola, MD, PhD
José Francisco García Llorente,
José Luis Fernández Sueiro, MD, PhD*
Luis Francisco Linares, MD
M Carmen de Castro, MD
Mireia Moreno, MD
Pedro Zarco, MD
Rafael Ariza, MD, PhD
Xenofon Baraliakos, MD, PhD
Eugenio de Miguel, MD, PhD

**In memoriam*

Please address correspondence to:
Prof. Loreto Carmona,
Instituto de Salud Musculoesquelética
Calle Ofelia Nieto, 10
28039 Madrid, Spain.
E-mail: loreto.carmona@inmusc.eu

Received on February 2, 2013; accepted in revised form on May 27, 2013.

Funding: this exercise was supported finantially by Abbott laboratories.

Competing interests: none declared.

Introduction

Spinal inflammation is one of the most important features of ankylosing spondylitis (AS), as it determines spinal pain and dysfunction (1). Currently, spinal inflammation is mainly assessed through patient reported outcomes (2, 3) and it may be subject to individual factors and interferences (4-6). Until the last decade, imaging in AS was confined to the use of x-ray, a technique which is able to assess structural damage solely. Various indices were developed, such as the BASRI or the mSASSS (7-10), which quantify the extent and degree of injury in terms of damage, not actual inflammation (8). Magnetic resonance imaging (MRI) has demonstrated its value in detecting bone oedema and inflammation and it has been included in the new classification criteria of the Assessment of Spondyloarthritis international Society (ASAS) (11). Gradually, MRI was imposed due to its greater sensitivity compared to x-rays, especially in early lesions and in clinical trials (12, 13). Although MRI has not entirely supplanted conventional radiology – among other reasons, due to its high cost and the time needed to perform and read – it is important to know how to evaluate MRI images in AS. This is especially relevant in order to conduct research and to understand the results of clinical studies, and also in order to be able to apply this knowledge to individual patients in whom a MRI is clinically justified.

There are several methods to score MRI lesions in AS, the most well-known being the Ankylosing Spondylitis Spine Magnetic Resonance Imaging (ASspiMRI), and its Berlin modification (Berlin score) (14), the Spondyloarthritis Research Consortium of Canada (SPARCC), and the Leeds methods (15). All of them are valid and comparable in terms of feasibility (16, 17).

The ASspiMRI evaluates activity and chronicity of the MRI lesions, while the Berlin score grades activity purely. In the latter, individual lesions are scored on a scale of 0 to 3 for every vertebral unit assessed. Disease activity is routinely assessed by using the short-tau inversion recovery (STIR) technique to visualise bone marrow oedema or T1-weighted MRI techniques after application of contrast agents. The Berlin score uses a semi-quantitative analysis of at least two images per lesion, and assesses all accessible vertebrae, from C2 to S1 (23 vertebral units), including the inter-vertebral disc space, in what is called a vertebral unit (18).

Both the ASspiMRI and the Berlin modification are validated instruments but with downsides, such as a reasonable doubt of reproducibility in clinical practice or a feasibility that may need improvement when used by inexperienced readers. Furthermore, there is indirect evidence of a moderate feasibility of the Berlin score, which is neither better nor worse than that of other MRI scoring methods (14, 17). Additionally, any imaging scoring method available is characterised by a high subjectivity and by change after increased reading experience (19). There are previous experiences of multi-reader experiment with the ASspiMRI that show a moderate to good inter-reader variability, but with the qualification that most studies have been done with readers trained by the developer of the technique (16, 18). The study of its reliability outside a clinical study and the effectiveness of a formal training in form of a calibration exercise, as well as of the factors that could improve its feasibility, are therefore justified in this context.

Hypothetically, the Berlin score may have a low intra-and interobserver reliability among untrained rheumatologists. A formal training in the Berlin score should decrease variability and improve the concordance with an expert reader. The main objective of this study was thus to evaluate the effectiveness of a formal training of inexperienced rheumatologists in reading spinal MRIs of AS patients with the Berlin score with the target of improving reliability and concordance with an expert reader. A secondary objective was to study the feasibility of the Berlin score in an environment outside a clinical study.

Methods

The calibration exercise consisted of a two-day theoretical and practical workshop on how to read and score MR im-

Table I. Results of the Berlin score assessment in 20 ankylosing spondylitis patients before and after a calibration exercise in 11 inexperienced readers.

	Berlin score related to training							
	Before calibration exercise				After calibration exercise			
Patient	Mean	SD	min	Max	Mean	SD	min	max
1	5.5	3.7	0	12	7.5	3.6	0	14
2	11.3	4.1	6	21	12.6	3.9	8	21
3	11.5	4.4	2	16	14.3	5.0	8	25
4	1.5	2.6	0	9	2.1	2.5	0	8
5	6.7	2.4	4	11	9.9	3.5	5	14
6	11.0	2.4	7	15	11.7	5.5	4	20
7	0.1	0.3	0	1	0.9	1.8	0	6
8	1.4	1.4	0	4	3.4	2.2	0	8
9	1.5	1.6	0	4	2.8	2.5	0	8
10	12.4	1.4	10	15	15.4	3.6	10	20
11	23.6	3.4	18	30	25.1	6.8	16	36
12	1.1	0.7	0	2	2.0	2.2	0	8
13	0.9	0.5	0	2	1.2	0.9	0	3
14	1.5	1.5	0	5	2.9	2.6	0	10
15	1.1	0.8	0	2	3.8	3.6	0	12
16	0.8	1.3	0	4	3.0	3.0	0	9
17	2.5	4.5	0	16	1.1	0.8	0	3
18	1.2	1.3	0	3	1.7	1.3	0	4
19	2.1	2.3	0	6	2.9	3.4	0	11
20	3.3	4.5	0	15	3.8	2.6	0	9
Total	5.0	6.4	0	30	6.4	7.1	0	36

Table II. Intra-reader reliability of the Berlin score by reader and calibration round.

	Intraclass correlation coefficient - Berlin score	
Reader	Before training	After training
1	0.87	0.83
2	0.92	0.82
3	0.99	0.96
4	0.81	0.79
5	0.78	0.85
6	0.95	0.94
7	0.83	0.76
8	0.97	0.96
9	0.96	0.81
10	0.85	0.85
11	0.87	0.98
Total*	0.89 (0.84–0.94)	.87 (0.82–0.92)

*Mean and 95% confidence interval.

ages by the Berlin score with one experienced reader (XB).

Thirteen rheumatologists with an interest in spondyloarthritis were invited to participate and read the images before and after the workshop. Reliability and concordance with the expert were compared before and two weeks after the training. Reliability was calculated from the reading scores of 20 patients with established AS who had complete sets of spinal MRIs. The selection of the images was performed for inclusion of all possible disease activity levels, from very high to very low. All MRIs were uploaded in an electronic database where any possible identification date was blinded. Every spinal MRI had at least 12 slices in the T1- and STIR-weighted sequence. The order of the patients presented in both sessions, before and after the training, was at random for each reader.

Intra-reader reliability was calculated from eight randomly selected MRIs, which were presented at random during the reading process. To avoid variability interfering with the reliability measure, the same readings were assigned at the same readers before and after the course.

The agreement with the expert assessment was calculated before and after the training from the scoring of the MRIs by the participating rheumatologists and by the expert. All the scores were blinded to the other readers.

Feasibility was measured as a discrete variable, based on the sum of the scores on two aspects: Total reading time (<2 min = 0, 2–5 min = 1, 5–8 min = 2, >8 min = 3) and complexity ("not complex at all" = 0, "somewhat complex" = 1, "fairly complex" = 2, "very complex" = 3). Items were collected while the reading was taking place. The total score range in feasibility was 0 to 6.

A web-based solution was developed to upload the 20 MRI exams, to gather the scores of the expert, and to allow reading and data collection. The web automatically registered the reader, date and time of start and end, for each scan. The data were downloaded to a spreadsheet for further statistical analyses.

The number of readings assigned to each rheumatologist was calculated based on the duration of the full exercise, which was not expected to exceed 5 hours in each repetition, and based on the expected variability, for a 20% detectable difference in reliability and concordance before and after.

The analyses conducted were:

1) Inter-reader reliability estimate before and after training, presented as the intraclass correlation coefficient (ICC) with 95% confidence interval (CI);

2) Intra-reader reliability before and after training, calculating ICC by reader and being the descriptive statistics for the exercise the mean of the coefficients at each exercise; and

3) agreement of the readings with those of the expert before and after training, measured again by the ICC and 95% CI.

All ICC were obtained from one-way analysis of variance with the patient as class variable. In addition, we studied the change in measurement between untrained readers and experts before and after the exercise by a Bland Altman graphic. The analysis of the feasibility of the method included a description of the measurement results and their components. All analyses were done with Stata 12.0 (StataCorp, College Station, TX).

Results

Of the 13 rheumatologists invited, the readings of two of them were excluded as either the first or the second readings were missing. Mean time to second reading after training was one month. Since each reader had scored 20 full patient images in each round, and repeated the scoring in 8, the total number of readings was 308 per round. The mean Berlin score (mean ± standard deviation; SD) was 5.0±6.4 in the first round and 6.4±7.1 after training ($p<0.01$). Table 1 shows the results of all cases before and after training.

Reading time per patient went from a mean ± SD of 7.2±5.4 minutes (minimum: 0.9; maximum: 38.5) in the first round to 5.7±3.6 (minimum: 0.5 maximum: 22.1) in the second round after training ($p<0.01$), which translates to a 20.8% time reduction. Regarding complexity, 117 readings (53%, excluding intra-reader scores) were considered very complex or fairly complex in the first round. In the second round this proportion fell to 103 (46%) ($p=0.182$). Feasibility score decreased from 3.3 ± 1.4 to 3.1 ± 1.3 ($p<0.01$).

Effectiveness of the training

Inter-reader reliability

In the first round, inter-reader ICC was 0.83 (95% CI: 0.75–0.93); in the second round ICC was 0.78 (95% CI: 0.66–0.90).

Intra-reader reliability

Before training the intra-reader reliability was ICC 0.89 (95% CI: 0.84–0.94); after training ICC was 0.87 (95% CI: 0.82–0.92). Table II shows individual intra-reader reliability of the 11 participants.

Agreement with the expert

The mean ± SD of the 20 scores by the expert were 15.0±11.2, practically 10 points on average over the trainees group. The agreement with the expert in the method was, before training, ICC = 0.59 (95% CI 0.45–0.76) and after training ICC = 0.65 (95% CI 0.50–0.80). Among the various readers, the agreement with the expert was moderately variable (Table III). Figure 1 shows the average between each pair reader-expert *versus* the difference between scores, before and after the exercise. The mean difference between the readings decreased after training. The graphic shows the slight and unbiased improvement.

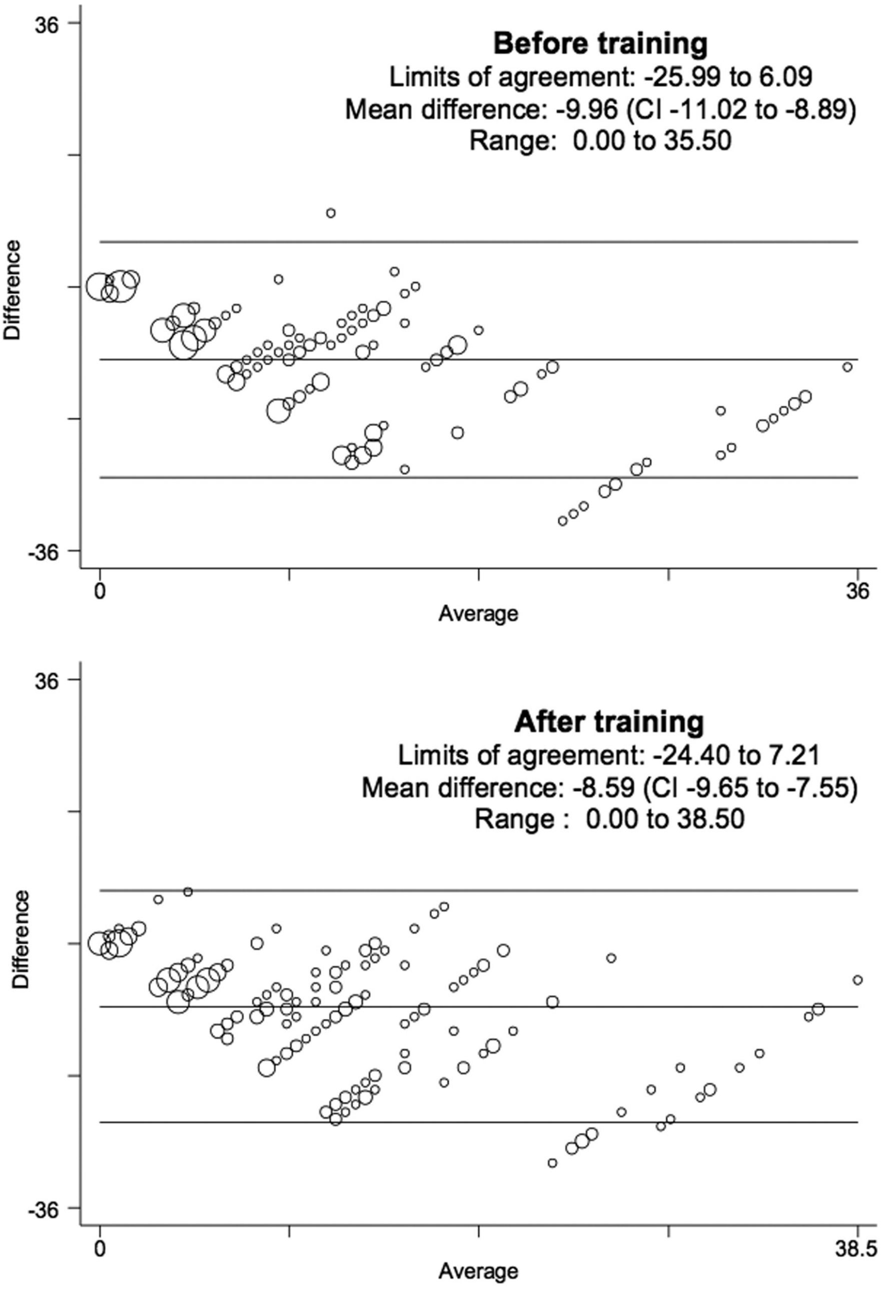

Fig. 1. Bland-Altman graphic comparing the Berlin scores of readers and expert before and after the training.

Footnote:

The x-axis contains the average between each score pair (the reading of an untrained reader and the expert's). These are plotted against the difference between both scores. The horizontal lines represent the mean difference plus the limits of agreement (defined as the mean difference plus and minus 1.96 times the standard deviation of the differences). The points on the Bland-Altman plot are scattered all over the graph, above and below zero, what suggests that there is no consistent bias of the readings by the expert *vs.* the inexperienced readers. There are fewer points outside the line of agreement in the second graph and the mean difference is smaller, what suggests that training has improved concordance.

Discussion

This experiment provides valuable information about the effectiveness of a calibration exercise on reading spinal MRIs in patients with AS in general and on the use of the Berlin score for quantification of these MRIs in particular.

One main deduction is that the calibration exercise increased the number of lesions detected by the initially inexperienced readers. The sum of scores, and the mean score in all patients but in one, increased 27% after the workshop. We could interpret this as if the training had increased the sensitivity of the rheumatologists in reading spinal MRIs of patients with active AS.

Table III. Concordance between the assessment of the expert and each reader, measured by intraclass correlation coefficient (95% confidence interval) before and after training.

	Concordance with the expert in Berlin score		
Reader	Before training	After training	Difference
1	0.23 (0.00–0.65)	0.51 (0.19–0.84)	0.29
2	0.22 (0.00–0.64)	0.43 (0.06–0.79)	0.21
3	0.25 (0.00–0.67)	0.38 (0.00–0.76)	0.13
4	0.25 (0.00–0.67)	0.42 (0.05–0.78)	0.17
5	0.41 (0.04–0.78)	0.17 (0.00–0.60)	-0.23
6	0.10 (0.00–0.54)	0.37 (0.00–0.75)	0.27
7	0.28 (0.00–0.69)	0.26 (0.00–0.67)	-0.02
8	0.43 (0.06–0.79)	0.70 (0.47–0.93)	0.27
9	0.33 (0.00–0.73)	0.35 (0.00–0.74)	0.01
10	0.10 (0.00–0.54)	0.34 (0.00–0.73)	0.24
11	0.27 (0.00–0.68)	0.17 (0.00–0.60)	-0.10
Mean	0.26	0.37	0.11

Reading after the training was also significantly faster (20%) and readers had a lesser perception of difficulty. Therefore, a second inference could be that the training improved the feasibility of the reading by the Berlin MRI scoring system.

With respect to the primary outcome of this reading exercise, improved reliability, we could not see much improvement. However this should be qualified, as the first readings already showed very high reliability, with intra and inter ICC over 0.80. This implies very little room for improvement with training, due to a ceiling effect. In addition, it called our attention that rheumatologists without a formal training in the method were able to score the MRIs with acceptable test-retest and inter-reader homogeneity, a finding that would support a modest need for training. Previous experiences training non-experts showed reliabilities after training as high as those shown before training in our study (10). In clinical trials and observational studies, reliability among experienced readers is not much higher than our baseline ICC (14, 18, 20, 21). This result actually speaks in favour of the stability and easiness of use of the Berlin MRI score.

Interestingly, despite good reliability, the initial concordance with the expert was if any only moderate but showed improvement after training. In comparison, in a multireader experiment of different scoring methods in which readers were experts, the agreement among readers was similar to what we found in our exercise (14). For the results presented in this exercise, concordance with the expert after training improved in the majority of the trainees, with an exception of four trainees whose scores dropped and intraobserver reliability worsened, probably in relation to loss of confidence. We assume that a longer training, and preferably with written feed-back, could increase the agreement with expert. However, the adequate length of training would need to be tested. In any case, a day and a half seems short, despite this is what usually workshops last, and this is what we wanted to test. Mixed face-to-face plus on-line training produce in general better results than any of them separately, but imply also a higher and longer commitment of experts.

In conclusion, the Berlin MRI score seems to be a reliable method even in the hands of inexperienced readers; calibration training may improve feasibility and increase the number of lesions detected but not to a large extent.

Acknowledgements

The authors want to acknowledge the support and enthusiasm by Beatriz Davalos, from Abbott, and the help from Let'sHealth with the web solution.

References

1. MACHADO P, LANDEWÉ R, BRAUN J, HERMANN KG, BAKER D, VAN DER HEIJDE D: Both structural damage and inflammation of the spine contribute to impairment of spinal mobility in patients with ankylosing spondylitis. *Ann Rheum Dis* 2010; 69: 1465-70.
2. ZOCHLING J: Measures of symptoms and disease status in ankylosing spondylitis: Ankylosing Spondylitis Disease Activity Score (ASDAS), Ankylosing Spondylitis Quality of Life Scale (ASQoL), Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), Bath Ankylosing Spondylitis Functional Index (BASFI), Bath Ankylosing Spondylitis Global Score (BAS-G), Bath Ankylosing Spondylitis Metrology Index (BASMI), Dougados Functional Index (DFI), and Health Assessment Questionnaire for the Spondylarthropathies (HAQ-S). *Arthritis Care Res* (Hoboken) 2011; 63 (Suppl. 11): S47-58.
3. PEDERSEN SJ, SORENSEN IJ, GARNERO P *et al.*: ASDAS, BASDAI and different treatment responses and their relation to biomarkers of inflammation, cartilage and bone turnover in patients with axial spondyloarthritis treated with TNF-alpha inhibitors. *Ann Rheum Dis* 2011; 70: 1375-81.
4. KILTZ U, BARALIAKOS X, KARAKOSTAS P *et al.*: The degree of spinal inflammation is similar in patients with axial spondyloarthritis who report high or low levels of disease activity: a cohort study. *Ann Rheum Dis* 2012; 71: 1207-11.
5. MACHADO P, VAN DER HEIJDE D: How to measure disease activity in axial spondyloarthritis? *Curr Opin Rheumatol* 2011; 23: 339-45.
6. ROUSSOU E, SULTANA S: Spondyloarthritis in women: differences in disease onset, clinical presentation, and Bath Ankylosing Spondylitis Disease Activity and Functional indices (BASDAI and BASFI) between men and women with spondyloarthritides. *Clin Rheumatol* 2011; 30: 121-7.
7. CASTREJON FERNANDEZ I, SANZ SANZ J: [Conventional Radiology: Total BASRI and SASSS]. *Reumatol Clin* 2010; 6S1: 33-6.
8. GURER G, BUTUN B, TUNCER T, UNUBOL AI: Comparison of radiological indices (SASSS, M-SASSS, BASRI-s, BASRI-t) in patients with ankylosing spondylitis. *Rheumatol Int* 2012; 32: 2069-74.
9. MACKAY K, MACK C, BROPHY S, CALIN A: The Bath Ankylosing Spondylitis Radiology Index (BASRI): a new, validated approach to disease assessment. *Arthritis Rheum* 1998; 41: 2263-70.
10. ULUSOY H, KAYA A, KAMANLI A, AKGOL G, OZGOCMEN S: Radiological scoring methods in ankylosing spondylitis: a comparison of the reliability of available methods. *Acta Reumatologica Portuguesa* 2010; 35: 170-5.
11. RUDWALEIT M, VAN DER HEIJDE D, LANDEWÉ R *et al.*: The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Ann Rheum Dis* 2009; 68: 777-83.
12. RUDWALEIT M, JURIK AG, HERMANN KG *et al.*: Defining active sacroiliitis on magnetic resonance imaging (MRI) for classification of axial spondyloarthritis: a consensual approach by the ASAS/OMERACT MRI group. *Ann Rheum Dis* 2009; 68: 1520-7.
13. BARALIAKOS X, LANDEWÉ R, HERMANN KG *et al.*: Inflammation in ankylosing spondylitis: a systematic description of the extent and frequency of acute spinal changes using magnetic resonance imaging. *Ann Rheum Dis* 2005; 64: 730-4.

14. LUKAS C, BRAUN J, VAN DER HEIJDE D *et al*.: Scoring inflammatory activity of the spine by magnetic resonance imaging in ankylosing spondylitis: a multireader experiment. *J Rheumatol* 2007; 34: 862-70.
15. VAN DER HEIJDE DM, LANDEWÉ 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.
16. BARALIAKOS X, HERMANN KG, LANDEWÉ 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.
17. VAN DER HEIJDE D, LANDEWÉ R, HERMANN KG *et al*.: Is there a preferred method for scoring activity of the spine by magnetic resonance imaging in ankylosing spondylitis? *J Rheumatol* 2007; 34: 871-3.
18. 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.
19. SHARP JT, WOLFE F, LASSERE M *et al*.: Variability of precision in scoring radiographic abnormalities in rheumatoid arthritis by experienced readers. *J Rheumatol* 2004; 31: 1062-72.
20. BARALIAKOS X, BRANDT J, LISTING J *et al*.: Outcome of patients with active ankylosing spondylitis after two years of therapy with etanercept: clinical and magnetic resonance imaging data. *Arthritis Rheum* 2005; 53: 856-63.
21. TAVONI AG, BALDINI C, BENCIVELLI W *et al*.: Minor salivary gland biopsy and Sjögren's syndrome: comparative analysis of biopsies among different Italian rheumatologic centers. *Clin Exp Rheumatol*. 2012; 30: 929-33.