OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 4: An International Multicenter Longitudinal Study Using the RA-MRI Score

PHILIP CONAGHAN, MARISSA LASSERE, MIKKEL ØSTERGAARD, CHARLES PETERFY, FIONA McQUEEN, PHILIP O'CONNOR, PAUL BIRD, BO EJBJERG, METTE KLARLUND, RON SHNIER, HARRY GENANT, PAUL EMERY, and JOHN EDMONDS

ABSTRACT. The aim of this multireader, multicenter study was to assess the inter-reader reliability of the score in the assessment of disease status and progression. The exercise involved 10 sets of metacarpophalangeal (MCP, 2nd to 5th) joints and 10 sets of wrist magnetic resonance images that were scored by experienced readers from 5 international centers. Synovitis was scored for each site using a global score (0–3). Bone abnormalities were assessed at 8 MCP joint sites and 15 wrist sites according to proportion of bone volume (0–10 for erosions and defects and 0–3 for edema). Intraclass correlation coefficients (ICC) and smallest detectable differences for synovitis, erosions, and edema were acceptable, although better for status scores than progression scores. The agreement for MCP joints was better than wrists. Limited variation in the images for some findings resulted in low ICC. Bone defects had the poorest agreement and have been omitted from new scoring recommendations. Despite limited training, multicenter readers demonstrated acceptable levels of agreement. (J Rheumatol 2003;30:1376–9)

Key Indexing Terms:
MAGNETIC RESONANCE IMAGING RHEUMATOID ARTHRITIS RELIABILITY

In rheumatoid arthritis (RA), magnetic resonance imaging (MRI) has the ability to image structural bone damage (erosions) as well as imaging measures of disease activity (such as synovitis). It can also image lesions that may reflect both disease activity and damage, known as bone edema. The OMERACT 5 RA-MRI working party previously reported the development of an RA-MRI scoring system (RAMRIS) for recording these MRI features[1]. Data have also been presented on the inter-reader reliability of this scoring method in cross-sectional patient cohorts, that is, when using the method to assess rheumatoid disease *status* at a single time point[2,3]. However, the reliability of this scoring method in assessing the change in measured variables over time (or *progression* score) has not been assessed.

The aim of the current study was therefore to examine inter-reader reliability of the RAMRIS scoring system when used to measure change over time in a longitudinal RA-MRI cohort set, as well as looking at the performance in measuring status.

From the Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, Leeds, UK; Department of Rheumatology, St. George Hospital, University of NSW, Sydney, Australia; The Danish Research Center of Magnetic Resonance and Departments of Rheumatology at the Copenhagen University Hospitals at Hvidovre, Herlev, and Rigshospitalet, Copenhagen, Denmark; Synarc, Inc. and Department of Radiology, University of California, San Francisco, California, USA; Department of Molecular Medicine, Auckland School of Medicine, University of Auckland, Auckland, New Zealand; Department of Radiology, Leeds General Infirmary, Leeds, UK; Mayne Nickless Sydney Imaging Group, Sydney, Australia; Rheumatology Unit, CHU Nantes, Nantes, France; Department of Radiology, University of California, San Francisco, California, USA.

P. Conaghan, MB, BS, FRACP, Senior Lecturer in Rheumatology, Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds; M. Lassere, MB, BS, Grad Dip Epi, PhD, FRACP, FAFPHM, Staff Specialist in Rheumatology, Senior Lecturer in Medicine, Department of Rheumatology, St. George Hospital; M. Østergaard, MD, PhD, DMSc, Professor of Rheumatology/Arthritis, Danish Research Centre of Magnetic Resonance and Departments of Rheumatology, Hvidovre, Herlev and Rigshospitalet; C. Peterfy, MD, PhD, Chief Medical Officer, Synarc, Inc. and Department of Radiology, University of California, San Francisco; F. McQueen, MD, FRACP, Senior Lecturer in Rheumatology, Department of Molecular Medicine, Auckland School of Medicine; P. O'Connor, MB, BS, MRCP, FRCR, Consultant Skeletal Radiologist, Department of Radiology, Leeds General Infirmary; P. Bird, BMed(Hons), Grad Cert MRI, FRACP, Research Fellow, Mayne Nickless Sydney Imaging Group; B. Ejbjerg, MD, Research Fellow, Synarc, Inc.; M. Klarlund, MD, PhD, Senior Registrar in Rheumatology, Rheumatology Unit, CHU Nantes; R. Shnier, MBBS, FRACR, National Director of Diagnostic Imagings, Consultant Radiologist, Mayne Nickless Sydney Imaging Group; H. Genant, MD, FACR, FRCR, Professor of Radiology, Medicine and Orthopaedics, Department of Radiology, University of California, San Francisco; P. Emery, MA, MD, FRCP, ARC Professor in Rheumatology, Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds; J. Edmonds, MB, BS, MA, FRACP, Director and Professor of Rheumatology, Department of Rheumatology, St. George Hospital.

Address reprint requests to Dr. P. Conaghan, Department of Rheumatology, Leeds General Infirmary, Great George Street, Leeds LS1 3EX United Kingdom. E-mail: Philip.conaghan@leedsth.nhs.uk

Methods

Design. The MR images of 10 pairs of RA dominant hand metacarpophalangeal (MCP) joints (from Sydney, Australia) and 10 pairs of RA dominant hand wrists (from Auckland, New Zealand) were scored separately by 5 and 4 readers, respectively. The readers were drawn from 5 international centers (Sydney, Australia; Leeds, UK; Hvidovre, Denmark; Auckland, New Zealand; San Francisco, USA) and had different levels of MRI reading experience. Scans were read from hard copy films (not workstations) in known time sequence.

All patients fulfilled the American College of Rheumatology 1987 criteria for RA. The MCP joint images were randomly selected from a RA cohort with established disease (median disease duration 6 yrs, interquartile range 10 yrs). The mean duration between first and second scans scored was 12 months. The wrist images were randomly selected from a cohort of early disease patients (duration < 12 mo at time of first scan) and the mean duration between first and second scans scored was 12 months. All patients were treated with disease modifying agents in the interscan time periods but no biologic agents were used.

MRI sequences. The MCP joints were imaged using a 1.5 T scanner (Siemens) with a dedicated hand coil. The field of view was 13 cm and included the 1st to 5th MCP joints. The imaging protocol comprised coronal (slice thickness 3 mm, no gap) and axial (slice thickness 4 mm, 1 mm gap) T1 sequences, followed by axial fat suppressed fast spin echo T2, then coronal fat suppressed T1 sequences after injection of gadolinium (Magnevist). The wrists were imaged using a 1.5 T scanner (GE Signa Horizon) with a dedicated wrist coil. The field of view was 8 cm and included the distal radioulnar, radiocarpal, and midcarpal joints as well as the metacarpal bases. The imaging protocol comprised coronal (slice thickness 3 mm, no gap) and axial (slice thickness 3 mm, 1 mm gap) T1 sequences, followed by axial fat suppressed fast spin echo T2, then coronal fat suppressed T1 sequences after injection of gadolinium (Nicomed Omniscan).

MRI scoring system. The definitions for synovitis, bone erosions, bone defects and bone edema have been described[1]. An erosion was defined as a bone defect with sharp margins, visible in 2 planes with a cortical break seen in at least one plane. A bone defect was defined as a sharply marginated area of trabecular loss without a visible cortical break. Bone edema could occur alone or surround a "defect" or "erosion," and was defined as a lesion with ill defined margins that was neither erosion nor defect and had high signal intensity on T2 weighted sequences. Synovitis was the area in the synovial compartment that showed enhancement of a thickness greater than the width of the joint capsule after gadolinium.

Synovitis was scored using a global score 0–3. The 2nd to 5th MCP joints were scored giving a summated score range for global synovitis from 0 to 12. In the wrist 3 sites were scored (the radioulnar, radiocarpal, and intercarpal-carpometacarpal joints) giving a summated score range for global synovitis from 0 to 9.

Bone erosions and defects were scored 0–10 by the estimated volume of the defect as a proportion of the "assessed bone volume" using 10% increments (giving an interval-like measure) judged on all available images. For long bones, the assessed bone volume was defined from the cortex of the articular surface (or its best estimated position if absent) to a depth of 1 cm. For the carpal bones, the assessed bone volume was defined as the whole bone. Bone edema was scored 0–3 by the volume of edema as a proportion of the assessed bone volume. The bone abnormalities were assigned to the proximal and distal half of each 2nd to 5th MCP joint, meaning 8 sites were scored. Therefore for an individual scan the summated scale range for MCP joint was 0–80 for bone erosions and defects and 0–24 for bone edema, where 0 represents no abnormality present. In the wrist, 15 sites were scored: the base of 1st to 5th MCP joint, the 8 carpal bones, the distal radius, and distal ulna. Therefore the summated scale range for an individual wrist was 0–150 for bone erosions and defects and 0–45 for bone edema.

Statistical methods. Descriptive statistics (mean, minimum, maximum, standard deviation, median, 25th and 75th percentiles) were calculated for each reader (or center) and for each joint region at both time points and for the progression scores. Two methods were employed for assessment of inter-reader reliability. The single measure fixed effects intraclass correlation coefficient (ICC) was calculated[4]. A limitation of the ICC statistic is that if there is limited variation in the subjects to be rated, then the ICC value will be quite low despite there being trivial differences between reader measurements. Further, ICC are not robust to the effects of outliers and can be affected considerably by a few large agreements. Another method of assessing agreement is the smallest detectable difference (SDD)[5,6], which is derived from the limits of agreement method[7]. It quantifies random error using an absolute metric and is expressed in the same scale of measurement as the studied score. The SDD was calculated for all the summated MRI scores using the residual error variance from repeated measures analysis of variance (repeated measures ANOVA). As well, the SDD were calculated as a percentage of the highest score achieved for that joint region. The statistical programs used were Stata 7.0, SPPS 6.0.

Results

A total of 10 sets of 2nd to 5th MCP joints and 10 sets of wrist joints of RA patients were scored by 5 readers and 4 readers/centers, respectively. Descriptive statistics (mean, minimum and maximum scores) for the progression scores by reader/center and per joint region are presented in Table 1.

Table 1. Mean values (minimum, maximum) for summated MRI change scores for metacarpophalangeal joints (MCPJ) and wrists by international center.

	AU	UK	DK	NZ	USA
MCPJ					
Synovitis global [0–12]	–0.9 (–5.0, 1.0)	–0.7 (–4.0, 1.0)	0.1 (–4.0, 3.0)	0.4 (–2.0, 4.0)	–0.7 (–3.0, 0)
Bone erosions [0–80]	1.3 (0, 5.0)	0.6 (–1.0, 5.0)	0.3 (0, 1.0)	0.7 (–1.0, 4.0)	1.1 (0, 7.0)
Bone defects [0–80]	0 (–1.0, 1.0)	0.3 (0, 1.0)	0.1 (0, 1.0)	0.3 (0, 1.0)	0 (0, 0)
Bone edema [0–24]	–0.5 (–5.0, 1.0)	0.9 (–3.0, 4.0)	0.1 (–1.0, 2.0)	1.2 (–1.0, 4.0)	–0.6 (–6.0, 0)
Wrists					
Synovitis global [0–9]	–0.4 (–3.0, 0)	0.4 (–1.0, 1.0)	0.4 (–1.0, 1.0)	1.3 (0, 2.0)	
Bone erosions [0–150]	8.6 (–1.0, 9.0)	9.9 (–1.0, 12.0)	2.2 (–1.0, 3.0)	5.2 (–1.0, 10.0)	
Bone defects [0–150]	0.7 (0, 1.0)	–0.5 (0, 0)	0.3 (0, 0)	0.1 (–1.0, 2.0)	
Bone edema [0–45]	–0.5 (–5.0, 1.0)	0.9 (–3.0, 11.0)	0.1 (–1.0, 0)	1.2 (–1.0, 7.0)	

AU: Australia; UK: United Kingdom; DK Denmark; NZ: New Zealand.

The fixed effects ICC and SDD results per joint region are presented in Table 2. The data are presented for assessment of disease status at both first and second MRI time points and for the progression, or change over time, score. The ICC for MCP joints were generally better than those for wrists, where there was a more limited range of abnormalities. Overall the agreement for status was better than for progression. Agreement for bone defects was particularly poor in all areas.

Discussion

This study looked at inter-reader agreement using multiple readers from different international centers and employing the RAMRIS scoring system. The aim was to assess reliability in measuring progression as well as status. Overall the agreement for MCP and wrist progression scores was satisfactory for measures of synovitis, bone erosions, and bone edema. Bone defects had generally poor levels of agreement.

These results should be seen in context. The status scores for the measured MRI abnormalities were better than in the first OMERACT exercises[2], reflecting a training element as a result of subsequent collaborator meetings. Even so, the readers had limited formal training exercises and mostly this concerned MCP joints. Overall the agreement for progression and status appeared better for the MCP joint than for wrists. However, it should be noted that the spectrum of disease abnormalities was narrower in the wrist group (from an early RA cohort) and this would have had the effect of lowering the ICC obtained. The agreement statistics for the wrist scoring were also probably not as good as in the third OMERACT exercise[3], and this may reflect a change in readers involved in the current study.

As well, there were 5 or 4 readers for each anatomical region in this exercise. Generally inter-reader agreement is presented for only 2 raters, and intra-reader ICC for such scoring systems should be very high. Indeed, there are data for individual readers from this group of collaborators giving ICC of > 0.9.[8] The SDD should also be seen in context of other rheumatological outcome measures. The SDD for the swollen joint account (68 joints) is 67%, for a pain visual analog scale is 53%, and for radiographic outcome measures the Larsen and Sharpe scores have reported SDD of 30% and 21%, respectively[6].

Table 2. Intraclass correlation coefficients (ICC) and smallest detectable difference (SDD) statistics for inter-reader scoring of both status and progression in the metacarpophalangeal joints (MCPJ) and wrists.

	Status						Progression		
	First Time Point			Second Time Point					
	ICC	SDD	SSD %	ICC	SDD	SSD %	ICC	SDD	SSD %
MCPJ									
Synovitis global [0–12]	0.89	3.4	28	0.82	3.3	36	0.39	3.5	51
Bone erosions [0–80]	0.78	7.9	24	0.85	7.3	24	0.60	3.2	46
Bone defects [0–80]	0.54	1.7	42	0.33	1.8	47	0.05	1.1	54
Bone edema [0–24]	0.89	3.0	27	0.72	4.8	32	0.11	4.4	63
Wrists									
Synovitis global [0–9]	0.74	2.5	28	0.68	2.7	30	0.46	2.5	62
Bone erosions [0–150]	0.15	8.0	42	0.45	16.8	35	0.55	14.9	37
Bone defects [0–150]	–0.18	3.4	85	0.03	2.8	70	–0.07	3.5	70
Bone edema [0–45]	0.08	8.7	51	0.56	9.2	27	0.45	9.5	37

With respect to individual scored elements of RAMRIS, bone edema did not demonstrate as good levels of agreement as synovitis or bone erosions. The scoring for edema had been modified subsequent to OMERACT Exercise 3 and reduced to a 0–3 scale, but this does not appear to have improved agreement. As a result of the iterative process involved in the development of RAMRIS, the working party recommended that scoring of bone edema revert to the 0–10 scale by volume of edema and that bone defect scoring be dropped from the RAMRIS system[9].

In summary, this exercise using the RAMRIS scoring method has demonstrated reasonable agreement levels for multicenter readers in assessing change in MRI synovitis, bone erosions, and bone edema. Further exercises are planned that will include specific reader training and using data sets with a wide range of abnormalities.

REFERENCES

1. Conaghan P, Edmonds J, Emery P, et al. Summary of OMERACT activities, current status, and plans. J Rheumatol 2001;28:1158-61.
2. Ostergaard M, Klarlund M, Lassere M, et al. Inter-reader agreement in the assessment of magnetic resonance images of rheumatoid arthritis wrist and finger joints — an international multicenter study. J Rheumatol 2001;28:1143-50.
3. Lassere M, McQueen F, Østergaard M, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Exercise 3: an international multicenter reliability study using the RA-MRI score (RAMRIS). J Rheumatol 2003;30:1366-75.
4. Shrout P, Fleiss J. Intra-class correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
5. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999; 26:731-39.
6. Lassere M, van der Heijde D, Johnson K, et al. The reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for the smallest detectable difference, the minimum clinical important difference and the analysis of treatment effects in randomised controlled trials. J Rheumatol 2001; 28:892-903.
7. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
8. Bird P, Ejbjerg B, McQueen, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Exercise 5: an international multicenter reliability study using computerised MRI erosion volume measurements. J Rheumatol 2003;30:1380-4.
9. Østergaard M, Peterfy C, Conaghan P, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system. J Rheuamtol 2003; 30:1385-6.