OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 3: An International Multicenter Reliability Study Using the RA-MRI Score

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

ABSTRACT. We examined inter-reader agreement of the revised OMERACT 5 Rheumatoid Arthritis MRI Score (RAMRIS v3). Magnetic resonance (MR) images of 10 sets of metacarpophalangeal (MCP) joints 2–5 and 8 sets of rheumatoid arthritis (RA) wrists [1.5 T, coronal and axial T1 and T2 spin-echo, ± fat saturation (FS), ± intravenous gadolinium (Gd)] were scored for (1) synovitis using a global score (0–3) and a direct measurement of synovial thickness (mm) and (2) three bone lesions: erosions, defects and edema, (score 0–10 by the volume of the lesion as a proportion of the "assessed bone volume" by 10% increments). Six readers from 5 multinational centers performed all scoring. Three statistical methods were used to analyze the data: (1) single-measure fixed effects intraclass correlations (sICC) and average-measure fixed effects ICC (avICC), (2) percentage exact and close agreement, and (3) the smallest detectable difference (SDD). The sICC were moderate to good (between 0.60 and 0.91) for half of the joint sites for the 2 synovitis scoring methods, and for bone erosions and bone edema. After adjusting for 6 readers, the avICC was very good to excellent (0.80–0.98) for two-thirds of the joint sites by lesion, excluding bone defects that performed relatively poorly, primarily because few readers scored these lesions. The aggregated scores with the best reliability were those with a wide range of scores, high ICC, low SDD, and low percentage SDD (< 33%). The metacarpophalangeal (MCP) bone erosion (sICC 0.58, avICC 0.89, %SDD ± 27), wrist bone erosion scores (0.72, 0.94, ± 31%), the wrist synovitis global (0.74, 0.94, ± 32%), and synovial maximal thickness (0.6, 0.94, ± 32%) met these conditions. MCP joint synovitis global (0.76, 0.95, ±35%), MCP joint bone edema (0.63, 0.91, ± 34%), and wrist bone edema (0.78, 0.95, ± 38%) performed marginally less well. Bone defects performed poorly (MCP joint 0.18, 0.46, ± 56%; wrist 0.06, 0.24, ± 55%). The revised OMERACT 5 RAMRIS has acceptable inter-reader reliability for measures of disease activity (synovitis global and bone edema scores) and damage (bone erosion score). Whether the score is sensitive to change will be determined by its performance in longitudinal and intervention studies. (J Rheumatol 2003;30:1366–75)

Key Indexing Terms:
MAGNETIC RESONANCE IMAGING RELIABILITY RHEUMATOID ARTHRITIS SMALLEST DETECTABLE DIFFERENCE

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

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

Address reprint requests to Dr. M.N. Lassere, Department of Rheumatology, St. George Hospital, Gray Street, Kogarah, 2217 NSW, Australia. E-mail: lasserem@sesahs.nsw.gov.au

Developing a magnetic resonance imaging (MRI) measurement system to evaluate the various facets of rheumatoid arthritis (RA) disease activity and damage is a multistep process. The following require consideration: which joints [all metacarpophalangeal (MCP) joints, selected wrist joints, dominant hand, etc.]; what anatomic structures or tissues (bone, cartilage, synovium, tendon, etc.); what features (bone edema, bone erosions, cartilage volume, synovial thickening, etc.); how to quantify these features (grade, score, count, or calculate areas or volume of anatomic structures and lesions); whether to weight this quantity and if so, how (implicitly, explicitly); whether and how to aggregate (Boolean or arithmetic operators) the quantified features into a component score. Moreover, at each step sources of variability, reliability, validity, responsiveness, and feasibility[1] should be evaluated in field trials using appropriate methods of statistical analysis.

However, many of the MRI measurement methods developed for use in RA have not been rigorously evaluated. A recent review of 68 peer reviewed publications of the MRI literature on measurement methods developed for RA, from 40 research groups[2], found that all (but one) reported the MRI variables, but only 35% of published studies were evaluated for reliability, and of these, few used optimal statistical methods of analysis. Responsiveness faired only slightly better (37% of studies). However, almost all studies (88%) evaluated some form of validity.

In 1998, after a meeting at OMERACT 4, a MRI working group of rheumatologists, radiologists, and clinical epidemiologists was established to begin a rigorous, prospective, transparent, and data-driven research process to develop and evaluate new MRI scoring methods for use in RA. In 1999 at several international meetings preliminary scoring methods were developed [Rheumatoid Arthritis MRI Score version 1 (RAMRISv1) and version 2 (RAMRISv2)], and their interreader reliability was tested (Study Exercises 1 and 2), presented at OMERACT 5, and published[3]. In light of the results, the group proposed a revised MRI scoring system (version 3) and recommended a standardized protocol for MR image acquisitions[4].

This article presents the results of Exercise 3, which continues the iterative process of testing the reliability of the RAMRIS. It was hoped that with further training, reader calibration, standardization of imaging protocols, and with more precise definitions of lesions and their measurement, we could improve the performance of the revised MRI Score (version 3).

MATERIALS AND METHODS

Design

MR images of 10 sets of 2nd to 5th MCP joints (from Leeds) and 8 sets of RA wrists (from Auckland) were scored in 5 centers in 5 different countries by readers with various levels of MRI experience. All MR sets were read between August and December 2000. The 5 centers were Auckland, New Zealand (FM, NS), Hvidovre, Denmark (MØ, MK), Leeds, United Kingdom (PC, POC, PE), San Francisco, USA (CP, HG), and Sydney, Australia (ML, PB, RS, JE).

Readers

There was no prespecification regarding the reading process. In Sydney, the 2 readers [PB, rheumatologist (Sydney 1) and RS, radiologist (Sydney 2)] independently read both sets. In Leeds 2 readers (POC, radiologist and PC, rheumatologist) read both sets, the results reflecting the consensus of both readers. In Copenhagen, 2 readers (MO and MK, rheumatologists) read both sets, the results reflecting the consensus of both readers. In Auckland one reader (FM, rheumatologist) independently read the MCP set, and 2 readers (FM and NS, radiologist) read the wrist set, the results reflecting the consensus of both readers. Finally, in San Francisco one reader (CP, radiologist) independently read both sets. In summary, there were 6 sets of results from 5 centers for both the MCP joint and wrist sets. All readers participated in the OMERACT 5 discussions and all but one reader took part in the previous exercises (Exercises 1 and 2). All statistical analyses were conducted with and without the results of this additional reader. There was no systematic difference depending whether 5 or 6 reader results were analyzed, therefore the results that include all 6 readers are provided in detail. However, the 5-reader results are available on request.

MRI Acquisition Specification and RA Patients

MR images of the 2nd to 5th MCP joints were obtained on a 1.5 Tesla MR scanner (Phillips). The field of view was 10 cm and was centered on the 2nd to 5th MCP joints. The imaging protocol comprised first, coronal (slice thickness 3 mm, no gap) and axial T1 weighted spin echo images, followed by coronal fat saturated T2 weighted images. After intravenous injection of gadolinium-DTPA contrast, the axial T1 weighted sequence was repeated, followed by a fat saturated coronal T1 weighted sequence. MR images of the dominant wrist were obtained using a 1.5 Tesla MR scanner (GE Signa Horizon) with a dedicated wrist coil (Medical Devices). 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 first, 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). The mean disease duration for the 10 MCP joint set and the 8 wrist joint set was 12 months. All patients fulfilled the American College of Rheumatology (ACR) 1987 criteria for RA.

Scoring of MRI

Structures included — definition of lesions. The revised OMERACT RAMRIS (version 3). An erosion was defined as a bone defect with sharp margins, visible in 2 planes (when 2 planes are available) 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. Cartilage was not scored because at Exercise 2 the demarcation of this tissue in the small joints of the wrist and hand was found to be too unreliable[3].

Scoring of bone lesions. A bone erosion lesion was scored from 0 to 10 by the *volume of the erosion* as a proportion of the "assessed bone volume" by 10% increments judged on all available images. For the carpal bones, the "assessed bone volume" was the whole bone. For long bones, the "assessed bone volume" was from the cortex of the articular surface (or its best estimated position if absent) to a depth of 1 cm. Bone defects and bone edema similarly were each scored 0–10 by the volume of the defect or of edema, as for erosion.

Bone erosions, bone defects, and bone edema were measured at 15 sites on the wrist image set and 8 sites (proximal and distal half of each 2nd to

5th MCP joint) on the MCP joint image set (see Figures 1 and 2 for scoring templates). Therefore the score for each MCP joint was 0–20, and the aggregated score for all 4 joints was 0–80 (where a score of zero indicates no erosions and a score of 80 indicates no bone). Bone defects and edema were scored similarly. In the wrist, 15 sites were scored: the base of 1st to 5th, the 8 carpal bones (hamate, capitate, trapezoid, trapezium, triquetrum, pisiform, lunate, scaphoid), and the distal radius and distal ulnar. Therefore the aggregated score for wrist was 0–150 for bone erosions, 0–150 for defects, and 0–150 for edema.

Scoring of synovitis. Synovitis was determined by gadolinium enhancement of the synovial compartment by 2 methods. In method 1, a global score of 0 to 3 was assigned, where 0 was normal with no synovial enhancement or enhancement no thicker than the joint capsule. Score of 1 to 3 was by thirds of the presumed maximum volume of enhancing tissue in the synovial compartment. This global score was assigned to the 4 MCP joint sites, giving an aggregated score of 0 to 12. In the wrist, the global score was assigned at 3 sites: the radioulnar joint; the radiocarpal joint; the intercarpal-carpometacarpal joints, giving an aggregated score of 0 to 9. In method 2, the maximum thickness of enhancing tissue in the slice showing the most thickening was directly measured in millimeters. This was measured at all 4 MCP joints on the axial view; in the wrists it was measured perpendicular to the cortical surface in the coronal view at the scaphoid and triquetrum; and in the axial view at the radioulnar joint and along the curved dorsal surface of the 1st and 2nd carpal rows.

Statistical Methods

The data were analyzed (1) individually by joint and lesion (i.e., synovitis global 2nd MCP, proximal 2nd MCP erosion, proximal 2nd MCP defect, etc.) to determine how agreement differed by joint and by lesion, and (2) as aggregated scores, analogous to the methods used to score radiographs, such as the Sharp radiographic score[5].

Descriptive statistics of each lesion (mean, minimum, maximum, standard deviation, median, 25th and 75th percentiles) for individual joints and aggregated scores were calculated by reader and across readers. Reliability was comprehensively evaluated with 3 statistical methods: intraclass correlation coefficient (ICC), percentage close/exact agreement, and smallest detectable difference (SDD). Three methods were used because each method entails certain assumptions that can produce biased results depending on the distribution of the scores under evaluation.

The first statistical method was the single measure fixed effects ICC (sICC) as described by Shrout and Fleiss[6]. ICC are a relative measure of agreement. The sICC and its 95% confidence interval[6] is similar to the quadratic weighted kappa for ordinal scale measures, where the weighted kappa is agreement beyond chance agreement. A second ICC, the average measure fixed effects ICC (avICC), corrects for the number of readers, so this was also provided[7].

One shortcoming of ICC is that if there is limited variation in the features scored, then the ICC value will be low despite trivial differences between reader scores. ICC values are biased towards high coefficients (1.0 is perfect reliability) if the data vary over a *wide* range. Another disadvan-

MRI Record Sheet – Metacarpophalangeal Joints (MCP)

Scoring Centre:_____ Scorer's name: _____ MRI ID:_____

Sequences scored: _____

Synovitis Scoring

	MCP			
	2	3	4	5
Synovitis - global score 0-3				
Synovitis – max thickness mm				

Bone Scoring

All bone abnormalities are scored 0-10 according to 10% volume involvement

NB score to depth of 1 cm from articular surface

		MCP			
		2	3	4	5
Bone erosion 0-10	Proximal				
	Distal				
Bone defect 0-10	Proximal				
	Distal				
Bone edema 0-10	Proximal				
	Distal				

Figure 1. MCP joint scoring template.

MRI Record Sheet – Wrist Joints (MCP)

Scoring Centre:_____ Scorer's name: _____ MRI ID:_____
Sequences scored: _____

Synovitis Scoring

Synovitis – global score 0-3	Radioulnar joint	Radiocarpal joint	Intercarpal-CMC

For measure scores, measure maximum enhancing tissue perpendicular to cortical surface

Synovitis – measure Coronal views Mm	Scaphoid	Triquetral

Synovitis – measure Axial views Mm	Radioulnar joint	Dorsal surface 1st/2nd carpal rows

Bone Scoring

All bone abnormalities are scored 0-10 according to 10% volume involvement

0-10	Base of Metacarpal				
	1	2	3	4	5
Bone erosion					
Bone defect					
Bone edema					

0-10	Hamate	Capitate	Trapezoid	Trapezium
Bone erosion				
Bone defect				
Bone edema				
0-10	Triquetrum	Pisiform	Lunate	Scaphoid
Bone erosion				
Bone defect				
Bone edema				

0-10	Distal Ulna	Distal Radius
Bone erosion		
Bone defect		
Bone edema		

Figure 2. Wrist scoring template. CMC: carpometacarpal.

tage of ICC is that they are not robust to the effects of outliers and can be affected considerably by a few large agreements.

To compensate for these problems we used a second statistical method, the percentage close agreement (PCA) and percentage exact agreement (PEA). PCA is based on a distribution of the degree of difference between readers. This is the percentage of occasions that different readers' scores fall within a certain distance of each other[8]. When calculating PCA, "close" is defined by judging what is meaningful for the measure and data set concerned. We set the PCA as within ± 1 interval for all lesions, including all aggregated lesions. Therefore percentage close agreement should be considered within the context of the actual score range.

The third statistical method was the smallest detectable difference[9,10], which is derived from the limits of agreement method[11]. Random error is quantified using an absolute metric. The SDD, unlike the ICC, is biased toward smaller values (SDD of 0 is perfect agreement, and there is no convention that anchors the upper limit) if the lesions are measured over a *narrower* range of values. The SDD is expressed in the *same units* of measurement as calculated for all aggregated scores. It is determined from the residual error variance of repeated measures analysis of variance (ANOVA)[12,13]. The SDD was also expressed as a percentage of the highest actual score to permit comparison of the reliability of MRI scoring with radiographic and clinical measures. Single factor repeated measures ANOVA was used to investigate whether "reader" was a significant source of variability. The statistical programs used were Stata 7.0[14], SPPS 6.0[7], and ICC.EXE[8].

RESULTS

Descriptive statistics (mean, range, standard deviations) of aggregated scores by lesion for each reader are shown in

Table 1. The full spectrum of the range was scored for the synovitis global lesions, particularly the MCP joints, less so the wrist joints, whereas bone erosions scored in the first quarter of the range for both MCP joints and wrists. A p value < 0.05 indicates that at least one reader differs from the others and that "reader" is a significant source of systematic variation in this dataset. Only for bone defects was the repeated measures ANOVA consistently not significant.

Averaging across readers (results not shown but available on request), the 2nd MCP joint consistently had the highest synovitis global score and the 4th MCP joint had the lowest. Repeated measures ANOVA did not show reader to be a significant source of variation for any MCP joint synovitis global score. The 2nd and 3rd MCP joints had higher scores for synovial maximal thickness, and reader was a significant source of variation only for 2nd MCP. The 2nd, 3rd, and 5th MCP consistently had higher erosion scores than the 4th MCP, and proximal scores were always greater than distal scores. Reader was a significant source of variability for 2nd and 5th MCP joints but not for 3rd and 4th MCP joints, irrespective of whether the joints were analyzed by their proximal or distal site. Defect scores were low for all joints, and there was no significant variability. The 2nd MCP joint consistently scored highest for bone edema, and proximal sites scored higher than distal sites. Reader was a significant source of variability for all but one site.

Average synovitis global scores at the 3 wrist sites did not differ, although synovial maximal thicknesses seen on coronal views were greater than those on axial. Reader variation was significant for all global scores and for 3 of the 4 maximal thickness scores. The 1st and 4th metacarpal bases, hamate, triquetrum, lunate, and scaphoid had the highest bone erosion and bone edema scores, followed by the capitate, trapezoid, trapezium, distal ulnar, distal radius, 2nd, 3rd and 5th metacarpal bases, and pisiform. There was little difference among all bones for "defects." Reader variation for bone erosions was significant for 1st metacarpal base, the capitate, trapezoid, triquetrum, lunate, scaphoid, distal ulnar, and radius; however, for bone edema it was significant at only the 1st metacarpal base, capitate, and triquetrum. It was not significant for defects.

The sites and lesions that performed best because joint/lesion showed a wide range of scores and because "reader" was not a significant source of variation on ANOVA were as follows: all MCP joints for synovitis global; 3rd MCP proximal site, 4th metacarpal base, and hamate for bone erosions; 4th metacarpal base, hamate, lunate and scaphoid for bone edema.

Table 2 shows the interreader fixed ICC and percentage agreement by joint/site and by lesion. Both the sICC and the avICC results are provided as well as the percentage exact agreement, or percentage close agreement within one interval. Bone defects usually scored zero at most sites, therefore as expected the ICC perform poorly and the percentage agreement was high. For the remaining joint/site by lesions the sICC was moderate to good (0.60–0.91) for half the joint sites by lesion. However, if the number of readers is taken into consideration, more than two-thirds of the joint sites by lesion (excluding bone defects) had very

Table 1. Means (minimum and maximum) of aggregated scores of MCP joints and wrists by different readers/centers.

	AU 1	AU 2	UK	DK	NZ	USA	ANOVA
MCP joints							
Synovitis global score (0–12)	4.7 (0–8)	6.5 (1–10)	6.7 (1–12)	6.6 (0–12)	5.7 (0–12)	6.7 (1–11)	0.027
Synovial maximal thickness, (mm)	11.1 (0–21)	12.9 (6–20)	11.9 (1–17)	13.2 (0–20)	9.8 (4–19)	8.6 (1–15)	0.020
Bone erosions proximal + distal (0–80)	8.2 (2–15)	10.4 (6–25)	5.8 (3–9)	4.7 (1–7)	4.7 (2–7)	8.3 (4–17)	0.000
Bone defects proximal + distal (0–80)	0.5 (0–2)	0 (0–0)	0.1 (0–1)	0.4 (0–2)	0.6 (0–3)	0 (0–0)	0.108
Bone edema proximal + distal (0–80)	3.2 (0–11)	7.9 (0–23)	11.3 (3–19)	3.3 (0–9)	4.5 (0–12)	4.3 (0–14)	0.000
Wrists							
Synovitis global score (0–12)	5.5 (2–9)	7.6 (5–9)	4.9 (3–9)	3.6 (0–7)	4.6 (2–6)	3.6 (2–5)	0.000
Synovial maximal thickness (mm)	21.3 (10–31)	23.4 (11–38)	9.8 (4–18)	16 (0–28)	9.8 (4–12.5)	19.9 (13–27)	0.000
Bone erosions (0–150)	13.5 (3–32)	17.1 (9–41)	13.8 (4–38)	2.8 (0–8)	6.6 (1–14)	15.8 (5–32)	0.000
Metacarpal bases (0–50)	3 (0–10)	4 (1–12)	3.8 (0–16)	1 (0-4)	1.8 (0–4)	4.3 (0–9)	0.021
Carpus (0–80)	8.1 (2–18)	11.4 (7–25)	9 (4–19)	1.5 (0–3)	3.8 (1–7)	9.3 (3–17)	0.000
Radius + ulnar (0–20)	2.4 (0–7)	1.8 (0–4)	1 (0–3)	0.3 (0–1)	1 (0–3)	2.3 (0–6)	0.000
Bone defects total (0–150)	1.5 (0–3)	0.25 (0–2)	0.13 (0–1)	1.6 (0–4)	2.8 (0–5)	0 (0–0)	0.000
Metacarpal bases (0–50)	0.3 (0–2)	0 (0)	0 (0)	0.6 (0–4)	0.9 (0–3)	0 (0)	0.090
Carpus (0–80)	1.3 (0–3)	0.3 (0–2)	0.1 (0–1)	0.9 (0–3)	1.3 (0–3)	0 (0)	0.011
Radius + ulnar (0–20)	0 (0)	0 (0)	0 (0)	0.1 (0–1)	0.3 (0–1)	0 (0)	0.220
Bone edema total (0–150)	10.13 (1–24)	10.2 (0–52)	24.3 (7–55)	14.1 (0–51)	10.9 (1–42)	8.6 (0–48)	0.002
Metacarpal bases (0–50)	2 (0–7)	2.9 (0–17)	8.1 (0–33)	4 (0–16)	2.9 (0–16)	2.9 (0–18)	0.056
Carpus (0–80)	6.5 (1–11)	6.5 (0–31)	14.1 (6–27)	9.4 (0–32)	6.9 (1–22)	4.8 (0–26)	0.015
Radius + ulnar (0–20)	1.6 (0–6)	0.7 (0–4)	2 (0–6)	0.8 (0–3)	1.1 (0–5)	1 (0–4)	0.099

Values in the 2nd to 7th columns are mean; max-min in parentheses. Values in 8th (right) column are p values. ANOVA: analysis of variance. AU 1: Australia Reader 1, AU 2: Australia reader 2; UK: United Kingdom, DK: Denmark, NZ: New Zealand, USA: United States of America.

Table 2. Interreader single and average measure intraclass correlation coefficient (ICC) and percentage agreement, selected by lesion and by joint/joint region. All ICC results are fixed effects.

	MCP JOINTS							
LESION	**MCP 2**		**MCP 3**		**MCP 4**		**MCP 5**	
Synovitis Global 0-3								
Single measure ICC	0.76		0.78		0.56		0.63	
Average measure ICC	0.95		0.95		0.88		0.91	
Percent exact agreement	63%		57%		50%		54%	
Synovial Max Thickness								
Single measure ICC	0.58		0.62		0.35		0.37	
Average measure ICC	0.89		0.91		0.77		0.78	
Percent close agreement	59%		60%		74%		78%	
	Proximal	**Distal**	**Proximal**	**Distal**	**Proximal**	**Distal**	**Proximal**	**Distal**
Bone erosion, 0-10								
Single measure ICC	0.62	0.13	0.71	0.46	0.62	0.09	0.57	0.62
Average measure ICC	0.91	0.48	0.94	0.84	0.91	0.38	0.89	0.91
Percent close agreement	78	92	85	97	84	96	NA	NA
Bone Defect, 0-10								
Single measure ICC	0.09	0.0	-0.02	NC	0.20	NC	-0.02	NC
Average measure ICC	0.37	0.0	-0.15	NC	0.60	NC	-0.15	NC
Percent close agreement	97	100	100	100	94	100	NA	NA
Bone Edema, 0-10								
Single measure ICC	0.77	0.78	0.36	0.31	0.11	0.09	0.25	0.66
Average measure ICC	0.95	0.96	0.77	0.73	0.41	0.37	0.67	0.92
Percent close agreement	56	88	71	93	98	97	NA	NA

	WRIST							
Synovitis Global, 0-3		**Radioulnar**	**Radiocarp**	**Intercarp**				
Single measure ICC		0.80	0.61	0.35				
Average measure ICC		0.96	0.90	0.76				
Percent exact agreement		44%	50%	33%				
Synovial Max Thickness		**Scaphoid**	**Triquetral**	**Radulnar**	**Dorsalcarpal**			
Single measure ICC		0.47	0.43	0.66	0.46			
Average measure ICC		0.84	0.82	0.91	0.81			
Percent close agreement		50%	33%	46%	61%			
	Metacarp1	**Metacarp 2**	**Metacarp 3**	**Metacarp 4**	**Metacarp 5**		**Distal Ulna**	**Distal Radius**
Bone erosion, 0-10								
Single measure ICC	0.70	0.13	0.47	0.43	0.91		0.73	0.72
Average measure ICC	0.93	0.46	0.84	0.82	0.98		0.94	0.93
Percent close agreement	83%	97%	96%	72%	76%		85%	86%
Bone Defect, 0-10								
Single measure ICC	NC	0.60	NC	0.45	NC		-0.21	NC
Average measure ICC	NC	0.75	NC	0.63	NC		-0.53	NC
Percent close agreement	100%	100%	100%	100%	95%		100%	100%
Bone Edema, 0-10								
Single measure ICC	0.60	0.09	0.90	0.93	-0.09		0.44	0.88
Average measure ICC	0.90	0.29	0.98	0.99	-0.19		0.82	0.98
Percent close agreement	76%	85%	90%	74%	76%		70%	92%
	Hamate	**Capitate**	**Trapezoid**	**Trapezium**	**Triquetrum**	**Pisiform**	**Lunate**	**Scaphoid**
Bone erosion, 0-10								
Single measure ICC	0.64	0.63	0.56	0.47	0.42	0.31	0.33	0.70
Average measure ICC	0.90	0.91	0.89	0.84	0.81	0.69	0.75	0.93
Percent close agreement	91%	82%	91%	83%	70%	100%	71%	90%
Bone Defect, 0-10								
Single measure ICC	-0.1 (3)	0.12 (3)	NC (0)	0.24 (3)	NC	-0.11 (2)	0.17 (3)	-0.11 (2)
Average measure ICC	-0.33	0.28	NC	0.48	NC	-0.26	0.38	-0.26
Percent close agreement	100%	100%	100%	100%	100%	100%	92%	100%
Bone Edema, 0-10								
Single measure ICC	0.54	0.90	0.20	0.61	0.29	0.10	0.23	0.52
Average measure ICC	0.88	0.98	0.56	0.89	0.71	0.31	0.64	0.87
Percent close agreement	74%	72%	89%	82%	58%	84%	58%	89%

PCA percentage close agreement within 1 interval. NC=not calculable because zero variance. NA = not analysed.
() number of readers whose scores were complete.

good to excellent avICC (0.80–0.98). The joint/site by lesions that remained unsatisfactory (avICC < 0.8 and percentage agreement < 80%) excluding bone defect scores were 4th and 5th MCP joint synovial maximal thickness, wrist intercarpal synovitis global score, 3rd and 5th MCP joint proximal, 5th metacarpal base, triquetrum, and lunate for bone edema, and only the lunate for bone erosions.

Table 3 shows the interreader fixed ICC and SDD statistics aggregated across sites by lesion. The aggregated scores that have the best reliability are those that show a wide range of scores, have a high avICC (> 0.80), low SDD, and low percentage SDD (< 33%). Usually SDD < 20% is preferred; however, this is less likely with 6 readers. The MCP bone erosion and wrist bone erosion scores, the wrist synovitis global and synovial maximal thickness, showed SDD below 33%. MCP joint synovitis global, MCP joint bone edema, and wrist bone edema scores had very good ICC, but the percentage SDD were just greater than 33%.

Finally, Table 4 summarizes the key results from Exercise 2 and compares these where applicable with Exercise 3 for the MCP joints. The MR image sets are identical, and the readers and sites, although not identical, are comparable between the 2 exercises (Exercise 3 had an additional reader). Several measures were modified or dropped between the 2 exercises. Synovitis global was scored in both exercises. Joint space narrowing and a 0 to 3 bone global score were dropped. The bone erosion score from Exercise 2 was modified from scoring 20% increments to score smaller increments of 10% of bone involvement; therefore the scoring scale was increased from 0–5 to 0–10. Exercise 2 bone lesion score combined bone defects and edema. These were separated in Exercise 3, and all were scored by 10% increments of involved bone. sICC results and SDD as a percentage of highest actual score are provided. Synovitis global scoring method was unchanged, and clearly the ICC and percentage SDD improved. Bone erosion scores show no improvement. "Bone lesion" in Exercise 2 was separated in Exercise 3 into "bone defects" and "bone edema." The score for bone edema improved by as much as the score for bone defect worsened.

DISCUSSION

Our study tested the inter-reader agreement of the revised rheumatoid arthritis MRI score developed by the OMERACT 5 MRI study group[3]. We found that with standardization of imaging protocols and more precise definitions of lesions and their measurement, the OMERACT RAMRIS (version 3) had acceptable reliability for 3 of the 5 lesions defined. Synovitis global, bone erosions, and bone edema demonstrated sICC of aggregated scores greater than 0.73 for most joints. A second method of assessing synovitis, the maximum thickness of enhancing tissue

Table 3. Interreader fixed intraclass correlation coefficient (ICC) and smallest detectable difference (SDD) statistics, selected by lesion and aggregated by joint/joint region and by readers, where % SDD is the smallest detectable difference as percentage of highest obtained score.

	Single Measure Fixed Effects ICC	Average Measure Fixed Effects ICC	SDD	SDD/Highest Actual Score, %
MCP joints aggregated				
Synovitis global, (0–12)	0.76	0.95	± 4.2	± 35
Synovial Max thickness (mm)	0.58	0.89	± 9.1	± 46
Bone erosion, (0–80)	0.51	0.86	± 6.6	± 27
Bone defect, (0–80)	0.18	0.47	± 1.8	± 56
Bone edema, (0–80)	0.63	0.91	± 7.7	± 34
Wrist regions				
Metacarpal bases aggregated				
Bone erosion, (0–50)	0.61	0.91		
Bone defect, (0–50)	0.33	0.59		
Bone edema, (0–50)	0.67	0.93		
Carpal row aggregated				
Bone erosion, (0–80)	0.63	0.91		
Bone defect, (0–80)	0.06	0.23		
Bone edema, (0–80)	0.62	0.91		
Radioulnar aggregated				
Bone erosion, (0–20)	0.70	0.93		
Bone defect, (0–20)	-0.21	-0.53		
Bone edema, (0–20)	0.75	0.95		
Wrist total aggregated				
Synovitis global, (0–9)	0.74	0.94	± 2.8	± 32
Synovial Max thickness (mm)	0.60	0.90	± 12.3	± 32
Bone erosion, (0–150)	0.72	0.94	± 12.6	± 31
Bone defect, (0–150)	0.06	0.24	± 2.8	± 55
Bone edema, (0–150)	0.78	0.95	± 20.6	± 38

Table 4. MCP joint: summary of exercises 2 and 3 aggregated joint scores.

Aggregated Scores	Exercise 2		Exercise 3	
	ICC	SDD/Highest Score, %	ICC	SDD/Highest Score, %
Image set		Leeds 10		Leeds 10
Reader		2 3 4 5 6		1 2 3ab 4ab 5 6
Synovitis			6	6
Global (0–3 per region)	0.59	43	0.77	35
Synovial maximal thickness	ND	ND	0.58	46
Joint space narrowing	0.25	90	ND	ND
Bone				
Global (0–3 per region)	0.36	37	ND	ND
Bone erosion (Exercise 2 by 20% increments) (Exercise 3 by 10% increments)	0.57	25	0.51	27
Bone lesion (combined bone defects and erosions, by 20% increments)	0.34	42	ND	ND
Bone defect	ND	ND	0.18	56
Bone edema	ND	ND	0.63	34

directly measured in millimeters, had marginally inferior reliability (ICC 0.58, 0.60). Only bone defects performed relatively poorly, in part because few readers scored bone defects, and this precluded valid evaluation of reliability using the intraclass correlation method.

The study design — requiring agreement across 6 readers with diverse levels of experience and background at 6 international sites without formal calibration — was a rigorous test of the revised MRI score. However, the design provided multifaceted and comprehensive data on several features of the score, valuable information for researchers and other users. Few RA imaging reliability studies have attempted to evaluate the reliability of a scoring method under these exacting conditions[2,10]. There are exceptions[15,16]. In 1985, Sharp, *et al*[16] evaluated the reliability of the Sharp radiographic scoring method across 9 readers without formal calibration. In the study by Fries, *et al*[15], 8 readers read practice films prior to the workshop preceding evaluation of 2 radiographic reading strategies being tested.

In our study we compared the sICC with the avICC, the latter adjusting for the number of readers[7]. The avICC improved the sICC by up to 0.3, the avICC now exceeding 0.9 for most lesions. To report an avICC requires that in future studies the mean score of multiple readers is used as the final score[6]. However, more than 2 readers are rarely used for any scoring method unless scoring is computerized[17,18]. Yet the discriminative capacity (precision) of measures is always improved by using the mean scores of multiple readers. The improved reliability can be calculated in advance using the Spearman-Brown prophecy statistic[13,15]. Fries, *et al*[15] found that using the mean score of 3 readers optimally improved the reliability of a radiographic scoring method, thereby improving study power and reducing costs in terms of sample size and operations in a hypothetical clinical trial.

Recently, the inter and intra-reader and inter-occasion agreement of the OMERACT 5 RAMRIS were evaluated on 12 MR wrist images. Intra-reader sICC exceeded 0.92 and inter-reader ICC exceeded 0.85 for synovitis global, bone erosion, and bone edema scores[19]. Studies of the reliability of other MR imaging protocols and scoring have demonstrated acceptable reliability[20-22].

Another difficulty with cross-study comparisons of reliability studies is that the results greatly depend on the data sets that are available and used for analysis. High ICC can be more easily achieved if the measure under assessment has at least some very low and some very high values so that interpatient variability is greater than interobserver variability[23]. The very low (and occasionally negative) ICC scores for bone defects reflected the almost universal "zero" scoring for this bone lesion. The percentage agreement was high, indicating that there was agreement among readers, but the higher than expected percentage SDD implies that agreement was indeed poor, even after considering the very poor spread of values.

Recently, the reliability of clinical, self-report physical function and quality of life measures, and radiographic scoring methods used as outcome measures in clinical trials were systematically reviewed[10]. The reliability of the RAMRIS in this study of synovitis global and bone erosions was equal to the reliability of most RA joint examination and self-report questionnaire assessment methods, and only slightly inferior to that for current radiographic scoring methods. This at the very least confirms the position of the reliability of MRI scoring for synovitis global and bone erosions as being comparable to that for most other

endpoints used in RA. Whether the OMERACT 5 RAMRIS meets the other elements of the OMERACT filter[1] — validity, responsiveness, and feasibility — is discussed elsewhere[24].

Our rationale for using 3 statistical methods to evaluate reliability was to ensure a comprehensive and accurate understanding. Determining and judging the reliability of a method is not a simple statistical undertaking, as we hope we have shown in this study. By using ANOVA based methods we have used statistics that require at least interval level of measurement[13]. The assessments of most radiographic scoring methods do the same. Although synovitis global score is ordinal, by aggregating across joints the score becomes more interval-like[25]. However, both bone erosion and bone edema scores, scored by the *volume of the lesion* as a proportion of the "assessed bone volume" in 10% increments, are interval-like measures prior to aggregation across joints. In Exercise 2[4], we defined 3 global methods of scoring lesions (synovitis, bone, and joint space narrowing), with bone erosions scored in 20% increments. Reliability was not evaluated for aggregated scores. Therefore, we used the kappa coefficient to ascertain agreement[26], and for selected measures the weighted kappa, equivalent to our sICC.

The MCP joint MRI image set was the same set used for OMERACT Exercise 2 (called substudy 2 in the publication), the first multicenter study to test inter-reader agreement on MR images of RA joints using an earlier MRI scoring method[4]. However, the wrist image sets from Exercise 2 did not meet the recommended imaging acquisition protocols, so a different wrist MR image set was used in Exercise 3, precluding direct comparison.

The reliability of scoring synovitis global clearly improved between exercises 2 and 3, suggesting that further training and more precise definitions of lesions were successful. However, reliability did not improve for the bone erosion score, and in fact overall it was marginally inferior. Perhaps reducing the incremental involvement of involved bone from 20% to 10% introduced more variability. Originally we planned to collapse the scores for bone lesions to 20% scaling (0–5) rather than use the scores as provided. This can be tested by further analysis of our data. Interestingly, "bone lesions," a term used in Exercise 2, was separated in Exercise 3 into "bone defects" and "bone edema", where the score for bone edema improved by as much as the score for bone defect worsened.

At 3 of the multicenter sites, the reader scores reflected the consensus readings of 2 experts. Scoring by consensus reduces variation. It is similar to taking the mean scores of 2 readers. In any inter-reader study the same mode of scoring should be employed, and future studies should specify whether reading occurs independently or by consensus. In Exercise 4, which evaluated the reliability of MRI change scores[27], all readers read the films independently. All but one of the readers in Exercise 3 participated in the previous OMERACT MRI exercises[4]. The new reader was a rheumatologist with 4 months of MRI scoring experience. Although he did not undergo any formal calibration by the MRI study group, this reader participated at all OMERACT 5 MRI study group meetings and trained under the guidance of an experienced reader who had participated in Exercise 2. To determine whether the addition of this new reader would influence the results, the data were analyzed with and without this reader's score. The final results did not differ, suggesting that dedicated readers can readily acquire expertise using the RAMRIS. It also illustrates that additional random variation incrementally decreases with the addition of readers.

Other sources of variability include the anatomy (e.g., the ligamentous insertion into capitate and the shape of the 4th MCP head), MRI factors (e.g., partial volume effects), lack of calibration, and lack of standard films. These are discussed by McQueen, *et al*[24].

As a consequence of Exercise 4, the OMERACT 5 RAMRIS system was further modified. The second method of determining synovitis, directly measuring the maximum thickness of enhancing tissue in millimeters, was not pursued, as much for reasons of parsimony and ease of scoring as for its performance. Also, bone edema scoring was revised from 10% increments to thirds (0–3 scale). The performance of these modifications as well as an evaluation of the reliability of the scoring method in longitudinal study are presented in Exercise 4[27].

In summary, we have shown in this study that the third revision of the RAMRIS has acceptable inter-reader reliability for measures of disease activity (synovitis global and bone edema scores) and disease damage (bone erosion score). Whether it is sensitive to change will need to be determined by its performance in longitudinal and intervention studies.

REFERENCES

1. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198-9.
2. Lassere M, Bird P. Measurement of RA disease activity and damage using MRI. Truth and discrimination: Does MRI make the grade? J Rheumatol 2001;28:1151-7.
3. Ostergaard M, Klarlund M, Lassere M, et al. Interreader 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.
4. Conaghan P, Edmonds J, Emery P, et al. Magnetic resonance imaging in rheumatoid arthritis: summary of OMERACT activities, current status, and plans. J Rheumatol 2001;28:1158-62.
5. Sharp J, Young D, Blubim G, et al. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis. Arthritis Rheum 1985; 28:1326-35.
6. Shrout P, Fleiss J. Intra-class correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
7. Statistical Package for the Social Sciences (SPSS) for Windows version 10. Chicago: SPSS Inc.; 1991-1999.

8. Chambers R, Adams R, Maher C. ICC.EXE: a program for calculating intraclass correlation coefficients and agreement indices. Sydney: School of Physiotherapy, Faculty of Health Sciences, University of Sydney.
9. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
10. Lassere M, van der Heijde D, Johnson K, Boers M, Edmonds J. 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 randomized controlled trials. J Rheumatol 2001;28:892-903.
11. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
12. Chinn S. The assessment of methods of measurement. Stat Med 1990;9:351-62.
13. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 1996.
14. Intercooled Stata 7.0 for Windows 2000. College Station, TX: Stata Corporation.
15. Fries JF, Bloch DA, Sharp JT, et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29:1-9.
16. Sharp JT, Bluhm GB, Brook A, et al. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. Arthritis Rheum 1985;28:16-24.
17. Sharp J, Gardner J, Bennett E. Computer-based methods for measuring joint space and estimating erosion volume in the finger and wrist joints of patients with rheumatoid arthritis. Arthritis Rheum 2000;43:1378-86.
18. Angwin J, Heald G, Lloyd A, Howland K, Davy M, James MF. Reliability and sensitivity of joint space measurements in hand radiographs using computerized image analysis. J Rheumatol 2001;28:1825-36.
19. Bird P, Lassere MN, Shnier R, Edmonds J. Computerised measurement of MRI erosion volumes in patients with rheumatoid arthritis — a comparison with existing MRI scoring systems and standard clinical outcome measures. Arthritis Rheum 2003; 48:614-24.
20. McQueen FM, Stewart N, Crabbe J, et al. Magnetic resonance imaging of the wrist in early rheumatoid arthritis reveals a high prevalence of erosions at four months after symptom onset. Ann Rheum Dis 1998;57:350-6.
21. Klarlund M, Ostergaard M, Gideon P, Sorensen K, Jensen KE, Lorenzen I. Wrist and finger joint MR imaging in rheumatoid arthritis. Acta Radiol 1999;40:400-9.
22. Huh YM, Suh JS, Jeong EK, et al. Role of the inflamed synovial volume of the wrist in defining remission of rheumatoid arthritis with gadolinium-enhanced 3D-SPGR MR imaging. J Magn Reson Imaging 1999;10:202-8.
23. Healy MJR. Measuring measuring errors. Stat Med 1989; 8:893-906.
24. McQueen FM, Lassere MN, Edmonds JP, et al. Rheumatoid arthritis magnetic resonance imaging studies. Summary of OMERACT 6 MRI module. J Rheumatol 2003;30:1387-92.
25. Hand DJ. Statistics and the theory of measurement. J R Statist Soc A 1996;159:445-92.
26. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
27. Conaghan P, Lassere MN, Ostergaard M, et al. Rheumatoid arthritis magnetic resonance imaging studies. Exercise 4: an international multicenter longitudinal study using the RA-MRI score (RAMRIS). J Rheumatol 2003;30:1376-9.

Papers presented at the OMERACT 6 Conference,
Gold Coast, Queensland, Australia, April 11–14, 2002.

Part 1: Patient Perspectives and Economics

Part 2: Imaging (Repair) and MCID/Low Disease Activity State

Part 3: Magnetic Resonance Imaging

Part 4: Outcome Measures for Clinical Trials: Systemic Sclerosis and Osteoarthritis

Part 4 will appear in the July issue of *The Journal*.